The Pervasiveness of Machine Learning in Omics Science

foundations, methods and applications

Ronnie Alves[1] , Claude Pasquier[2] & Nicolas Pasquier[3]

alvesrco@gmail.com

https://sites.google.com/site/alvesrco/tutorial_ecml_pkdd2014

[1]Institute of Computational Biology (IBC), LIRMM, Université Montpellier 2, France
[2]National Center for Scientific Research (CNRS)
[3]I3S Laboratory of Computer Science, Signals and Systems of Sophia Antipolis, Université Nice Sophia-Antipolis, France

ECML-PKDD'14, Tutorial T3 (room 105), 15/09/2014, Nancy

Pervasiveness

per.va.sive adj.

Noun. **pervasiveness** - *the quality of filling or spreading throughout*; *the quality of being general or widespread or having general applicability.*

http://www.thefreedictionary.com/ pervasiveness

Omics Science

Omics and Omes

"Omics is a general term for a broad discipline of science and engineering for analyzing the interactions of biological information objects in various "omes".

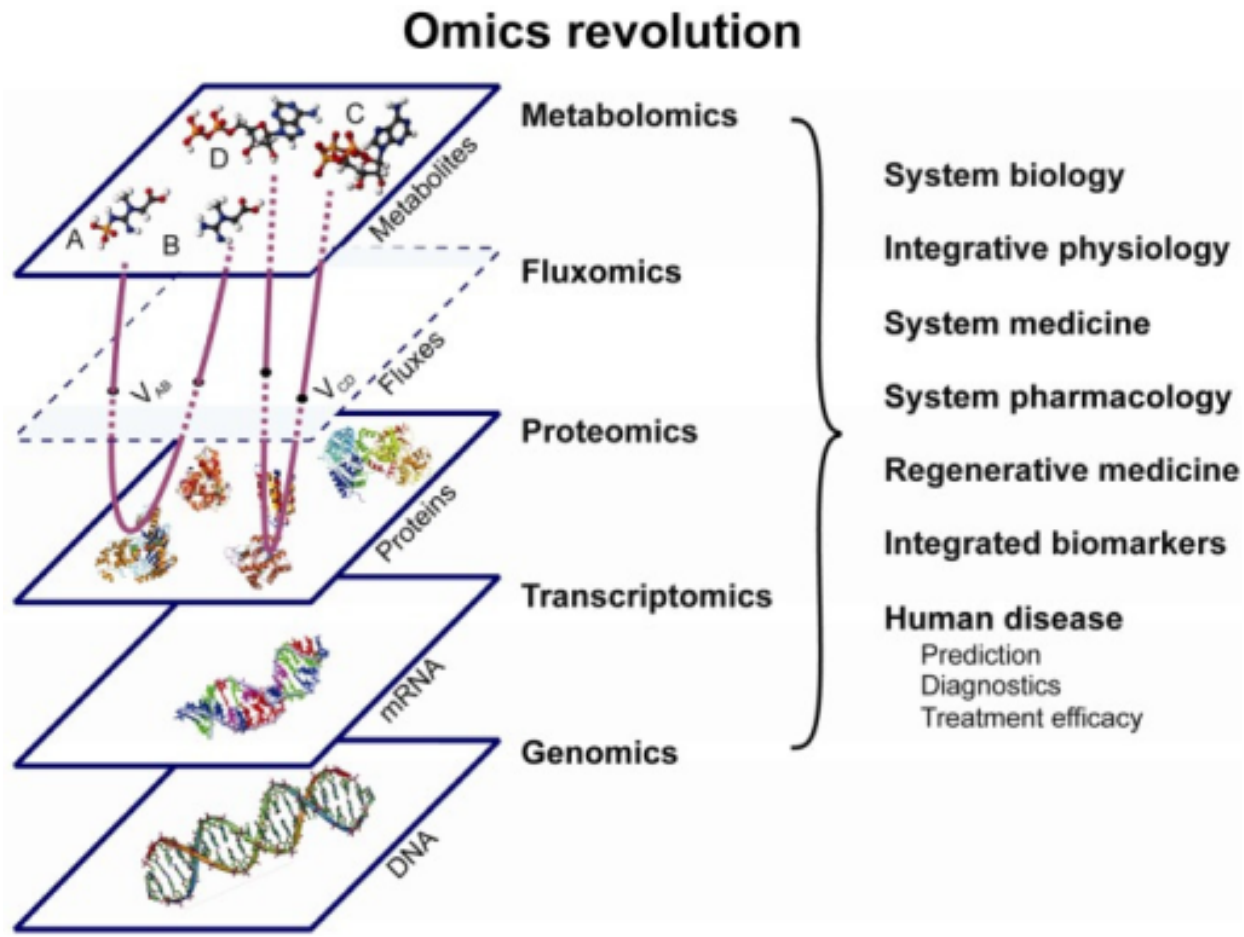
http://www.nature.com/omics/about/index.html
http://omics.org/index.php/Omes_and_Omics

Omics Science

The main focus is on

- mapping information objects such as genes, proteins, and ligands;
- finding interaction relationships among the objects;
- engineering the networks and objects to understand and manipulate the regulatory mechanisms; and
- integrating various omes and omics subfields.

Omics Science



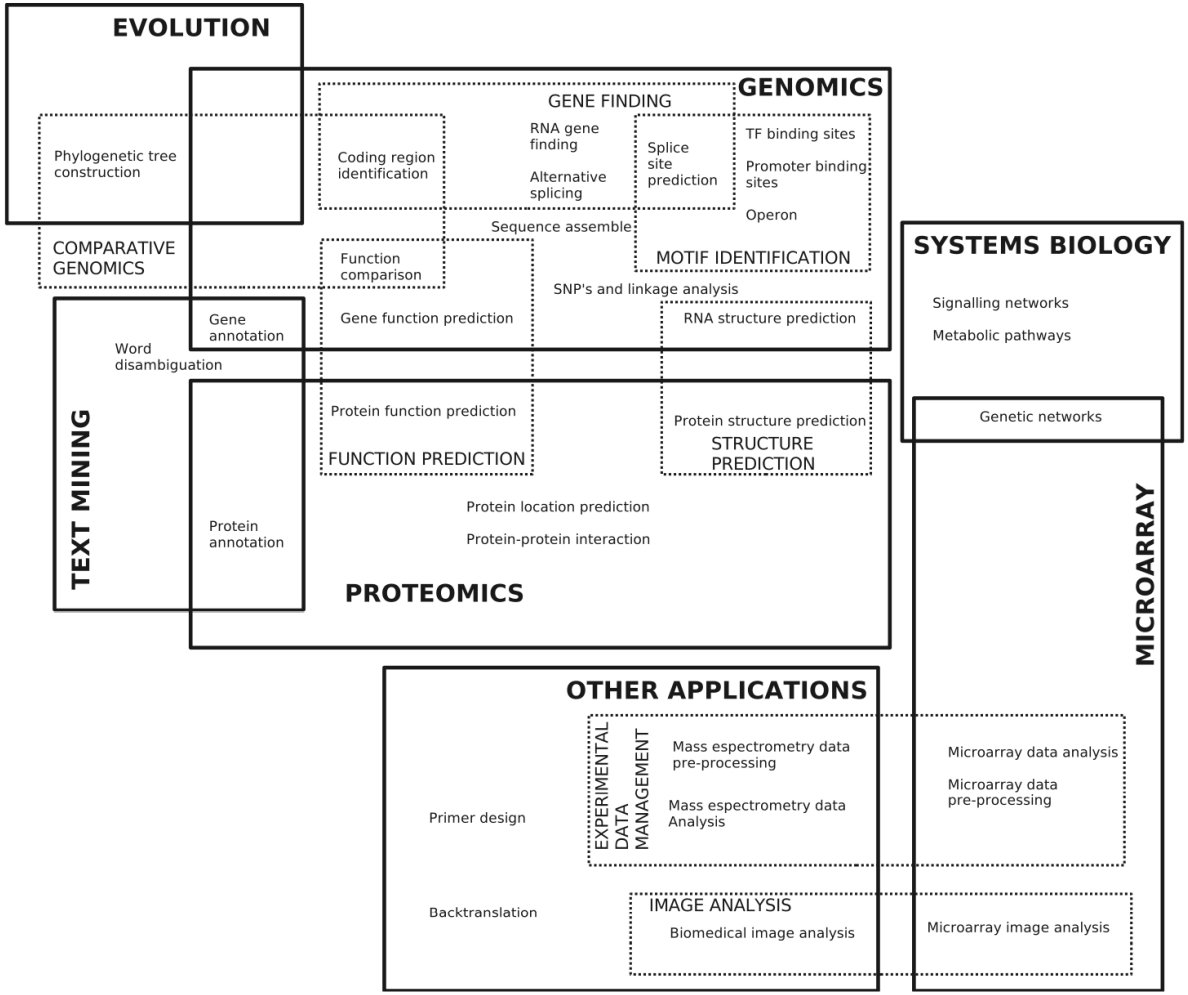
- "Omics" technologies will improve disease diagnostics and treatment by targeting drugs and procedures for each unique Omics profiles.

Biology has become a data-rich subject

Machine Learning (ML) in Omics

The exponential growth of the amount of biological data available raises basically two problems: i) efficient information storage and management and, ii) the extraction of useful information from these data. The latter is one of the main challenges in computational biology [4].

Biology has become a data-rich subject

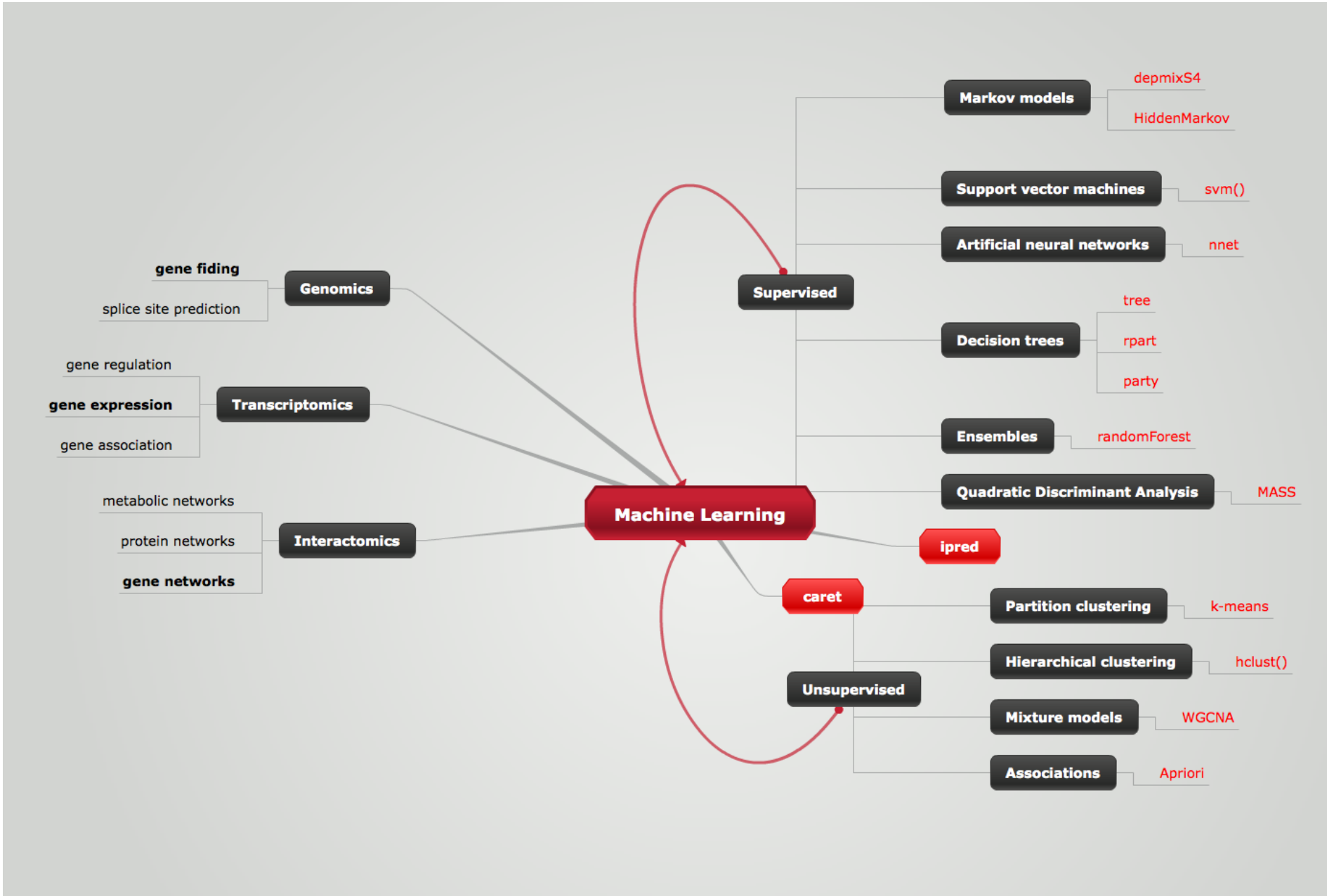


- An classification scheme of the topics where ML methods are applied [4].

Biology has become a data-rich subject

The ingredient of machine learning

"Models lend the machine learning field diversity, but tasks and features give it unity." [2]
Peter Flach's ML book.


Machine Learning
Supervised
Markov models
depmixS4
HiddenMarkov
Support vector machines
svm()
Artificial neural networks
nnet
Decision trees
tree
rpart
party
Ensembles
randomForest
Quadratic Discriminant Analysis
MASS
ipred
caret
Unsupervised
Partition clustering
k-means
Hierarchical clustering
hclust()
Mixture models
WGCNA
Associations
Apriori
Genomics
gene fiding
splice site prediction
Transcriptomics
gene regulation
gene expression
gene association
Interactomics
metabolic networks
protein networks
gene networks

Machine learning in genomics data

Data generation vs analysis

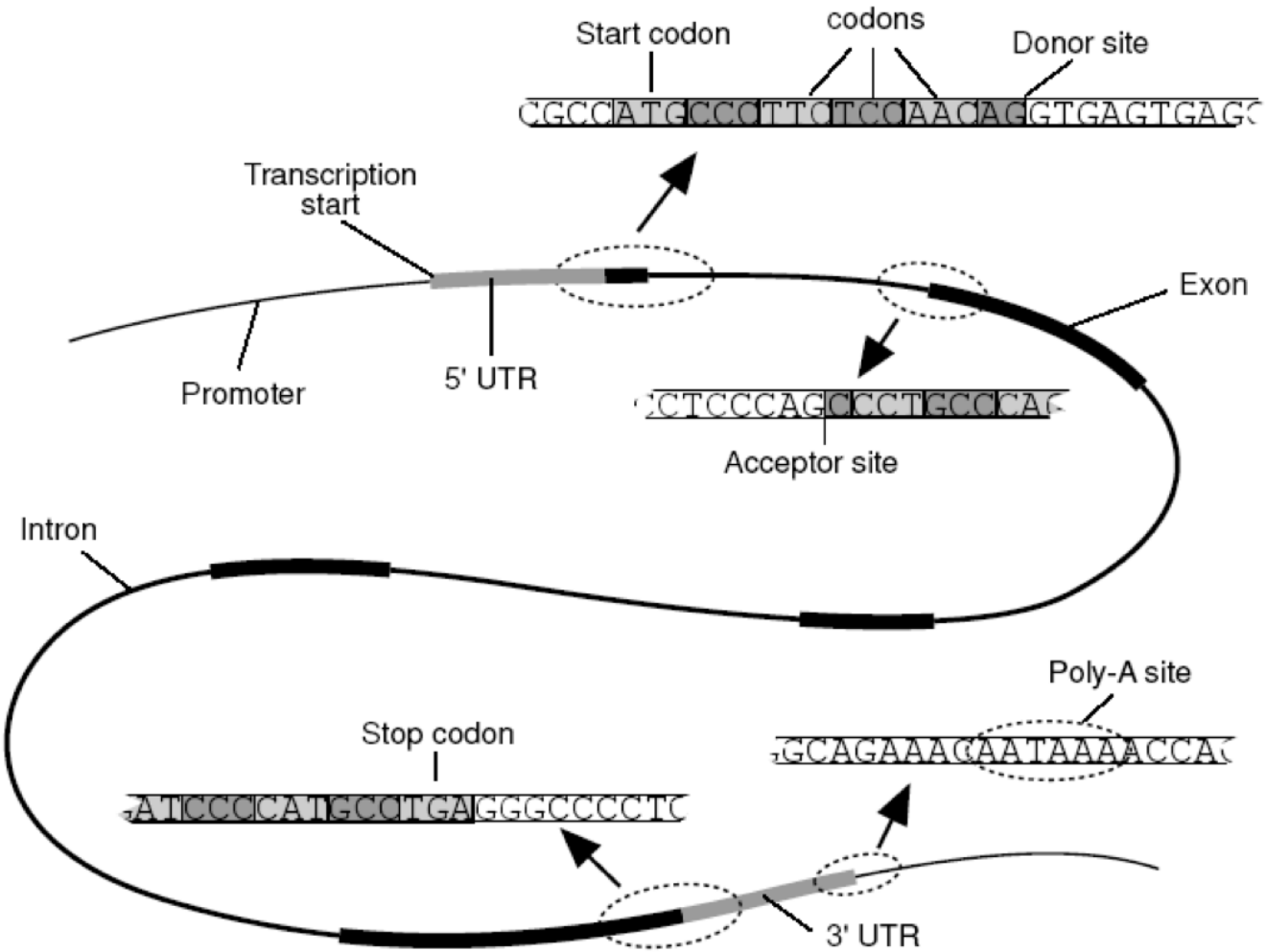
Bioinformatics needs

- traditionally, the U.K. and the U.S. have not invested in analysis; instead, the focus has been investing in data generation.
- analysis, not sequencing, will be the main expense hurdle to many genome projects.
- bioinformaticists are in short supply.

A general task

The gene finding problem

How to locate the genes along a genome?



A general task

The gene finding problem

How to locate the genes along a genome?

- Essentially, two different types of information are currently used to try to locate genes in a genomic sequence: i) content sensors and ii) signal sensors.

The gene finding problem

Extrinsic content sensors

Exploit a sufficient similarity between a genomic sequence region and a protein or DNA sequence present in a database in order to determine whether the region is transcribed and/or coding.

- Uses local alignment methods ranging from the optimal Smith-Waterman algorithm to fast heuristic approaches such as FASTA and BLAST.

The gene finding problem

Intrinsic content sensors

Originally, they were defined for prokaryotic genomes. Only two types of regions are usually considered: the regions that code for a protein and intergenic regions.

- The simplest approach for finding potential coding sequences is to look for sufficiently long open reading frames (ORFs), defined as sequences not containing stops, i.e. as sequences between a start and a stop codon.

The gene finding problem

What characterise a **coding sequence**?

i) Nucleotide composition and especially $(G + C)$ content[a], ii) codon composition, iii) hexamer frequency, iv) base occurrence periodicity, etc.

[a] introns being more A/T-rich than exons, especially in plants

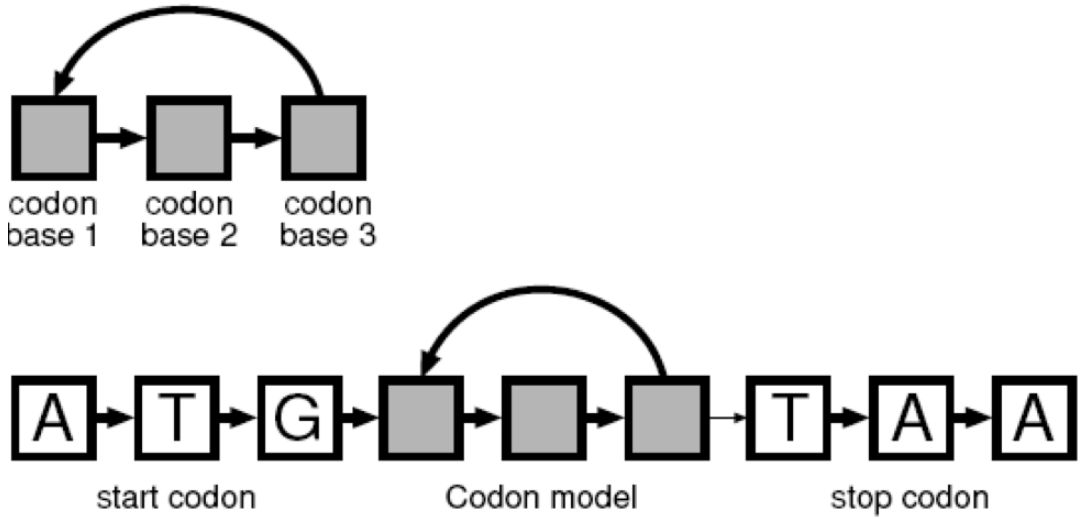
- The **hexamer usage** (i.e. usage of 6 nt long words) has been widely exploited by a large number of gene prediction algorithms through different methods.

The gene finding problem

Hidden Markov Models (HMMs): The "simple" model

HMM describes a **gene sequence** by a set of **states** (exons, introns, intergenic regions, splice sites, start and stop codons, etc.). Each state is characterised by emission probabilities and transition probabilities.

- **emission probability**: probability to observe an A in an exon sequence.
- **transition probability**: probability to have a splice site after an exon.



The gene finding problem

Hidden Markov Models (HMMs)

Build a model that uses sequence information from the previous five instead of the previous two bases to make what is called a **fifth-order Markov model**.

- Analyses of sequential codons in genes have shown that some pairs are found at a greater frequency and others at a lesser frequency than expected by chance alone.
- The frequency of **hexamers** is used to differentiate between coding and noncoding sequences.

The gene finding problem

GeneMark

It relies on **Inhomogeneous Markov Chain** (IMC) models.

- The Markov chains used are 5th-order, and consists of terms such as $P(a|b1b2b3b4b5)$ which represent the probability of the sixth base being a given the previous 5 bases $b1 - b5$.
- These probabilities must be defined for all possible pentamers with the general sequence $b1b2b3b4b5$. The values of these terms can be obtained by analysis of **training data**, consisting of nucleotide sequences in which the coding regions have been accurately defined.

The gene finding problem

GeneMark

Probability table (built from a known sequence)

5 bases	6th ba.	Prob. 6th in F1	Prob. 6th in F2	...
AAAAA	A	$P_1(A\|AAAAA)$	$P_2(A\|AAAAA)$	
AAAAA	C	$P_1(C\|AAAAA)$	$P_2(C\|AAAAA)$	
AAAAA	G	$P_1(G\|AAAAA)$	$P_2(G\|AAAAA)$	
AAAAA	T	$P_1(T\|AAAAA)$	$P_2(T\|AAAAA)$	
AAAAC	A	$P_1(A\|AAAAC)$	$P_2(A\|AAAAC)$	
...				
AAAAC	T	$P_1(T\|AAAAC)$	$P_2(T\|AAAAC)$	

The gene finding problem

GeneMark

GATCTAGCGTCA...

Probability of the 6th base given the previous 5 bases:

$$P(6th|b_1 b_2 b_3 b_4 b_5) = \frac{Nb_1 b_2 b_3 b_4 b_5 6th}{\sum_{i=A,C,GT} Nb_1 b_2 b_3 b_4 b_5 i}$$

where $Nb_1 b_2 b_3 b_4 b_5 6th$ is the frequency of the sequence in the coding regions.

The gene finding problem

GeneMark

The probability of obtaining the sequence $x = x_1x_2x_3x_4x_5x_6x_7x_8$ in a coding region in frame 3 is given by

$$P(x|3) = P_3(x_1x_2x_3x_4x_5)P_3(x_6|x_1x_2x_3x_4x_5)P_1(x_7|x_2x_3x_4x_5x_6)P_2(x_8|x_3x_4x_5x_6x_7)$$

Prob. in F3 $*$ Prob. in F3 $*$ Prob. F1 $*$ Prob. F2

The likelihood that the sequence x is in a coding frame 3

$$P(3|x) = \frac{P(x|3)P(3)}{P(x)} = \frac{P(x|3)P(3)}{P(x|nc)P(nc)+\sum_{m=1}^{6} P(x|m)P(m)}$$

GeneMark requires the compution of $4^6 = 4096$ possible di-codons probabilities + probabilities of occurrence of each pentamers.

The gene finding problem

GLIMMER

It circumvents the use of such a large number of probabilities by using a combination of Markov chains with various orders (**Interpolated Markov Models**)

- GLIMMER requires a training data, which is usually selected among known genes, genes coding for proteins with strong database hits, and/or simply long ORFs (typically longer than 500 bp).
- Using this training set, GLIMMER generates parameters for Markov models of increasing order from 0 to 8. For every order, a check is made if the number of observations of each sub-sequence is sufficient and, if not, the order is decreased.

It has been applied to the annotation of numerous microbial genomes.

The gene finding problem

GRAIL

It uses a **neural network strategy** to identify exons, polyA sites, promoters, CpG islands, repetitive elements, and frameshift errors in DNA sequences by comparing them to a database of known human and mouse sequence elements.

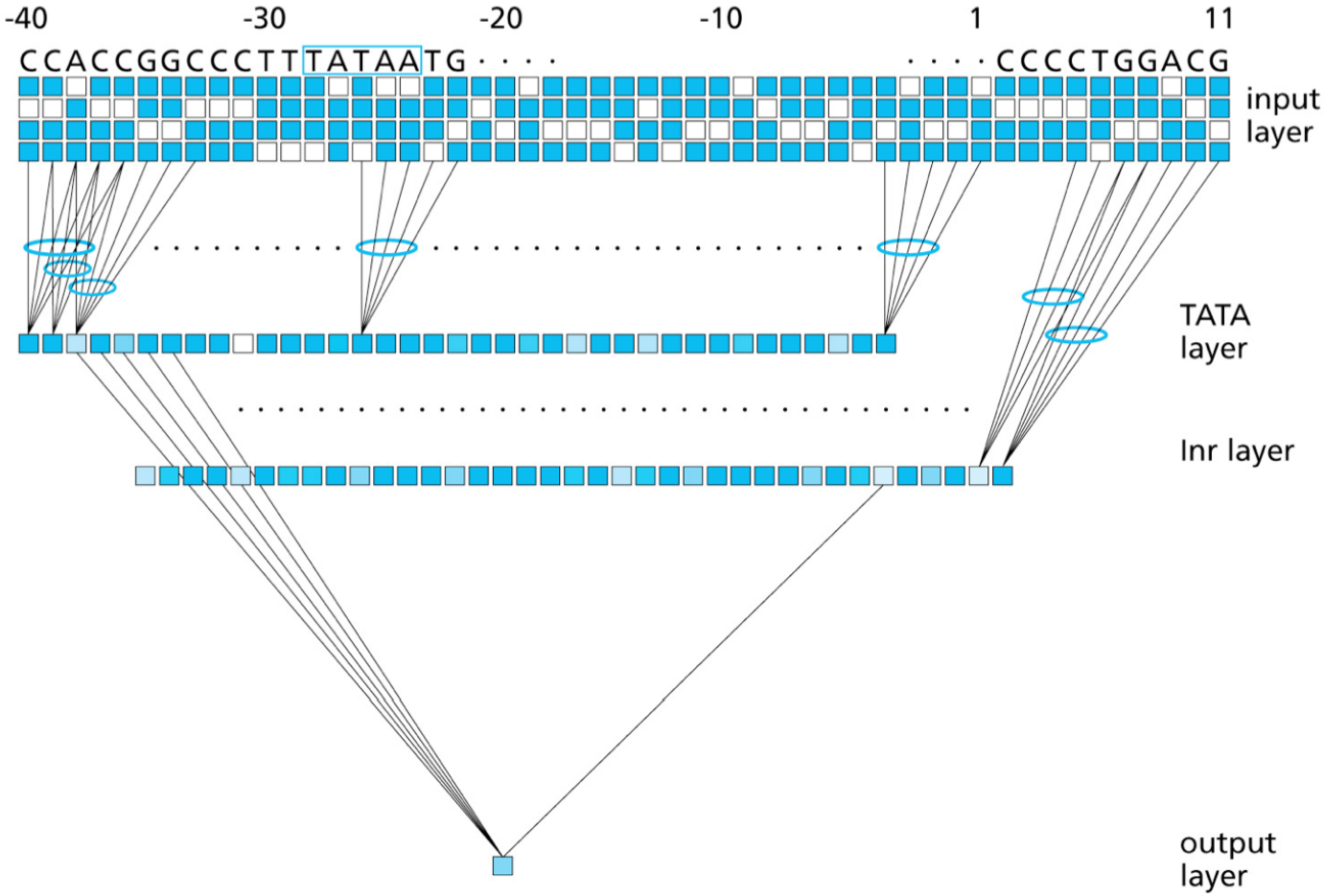
- **procaryotes**: GLIMMER + BLASTP(search predicted ORFs in Genbank and SWISS-PROT)
- **human and mouse**: GrailEXP + GenScan + BLASTP(search predicted ORFs in Genbank and SWISS-PROT)

Exon and repetitive element prediction is also available for *Arabidopsis* and *Drosophila* sequences.

The gene finding problem

GRAIL

A **neural network strategy** to identify TATA box and Inr signals.



The gene finding problem

MORGAN

It uses a **decision tree classifier** to identify start and stop codons, donor sites, acceptor sites and they are brought together in a frame-sensitive dynamic programming algorithm that finds the optimal segmentation (coding vs non-coding sequence).

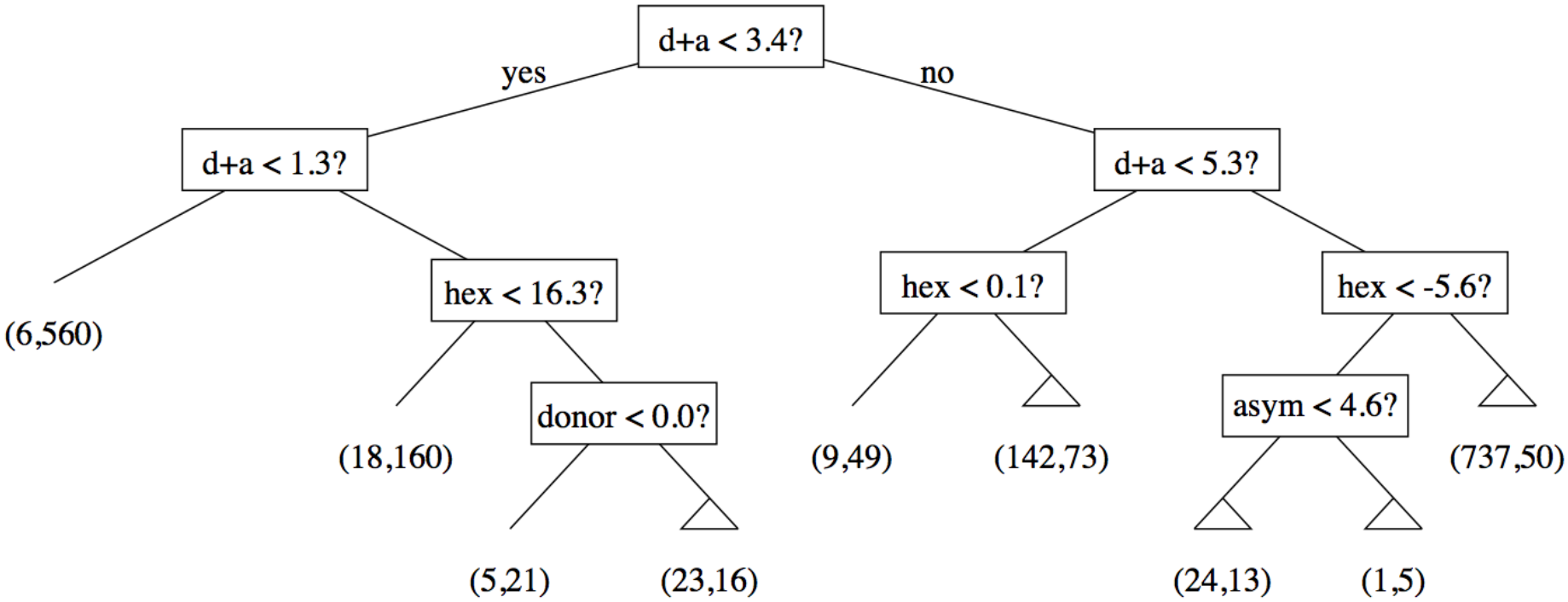
- The optimal segmentation is dependent on a separate scoring function that takes a subsequence and assigns to it a score reflecting the probability that the sequence is an **exon**.
- The **scoring functions** in MORGAN are sets of (randomised) decision trees that are combined to give a probability estimate.

Features: the start site score as computed by the conditional probability matrix, the donor and acceptor site scores from the Markov model the in-frame hexamer statistic and the position asymmetry statistic.

The gene finding problem

MORGAN

Sample decision tree for classifying human DNA.

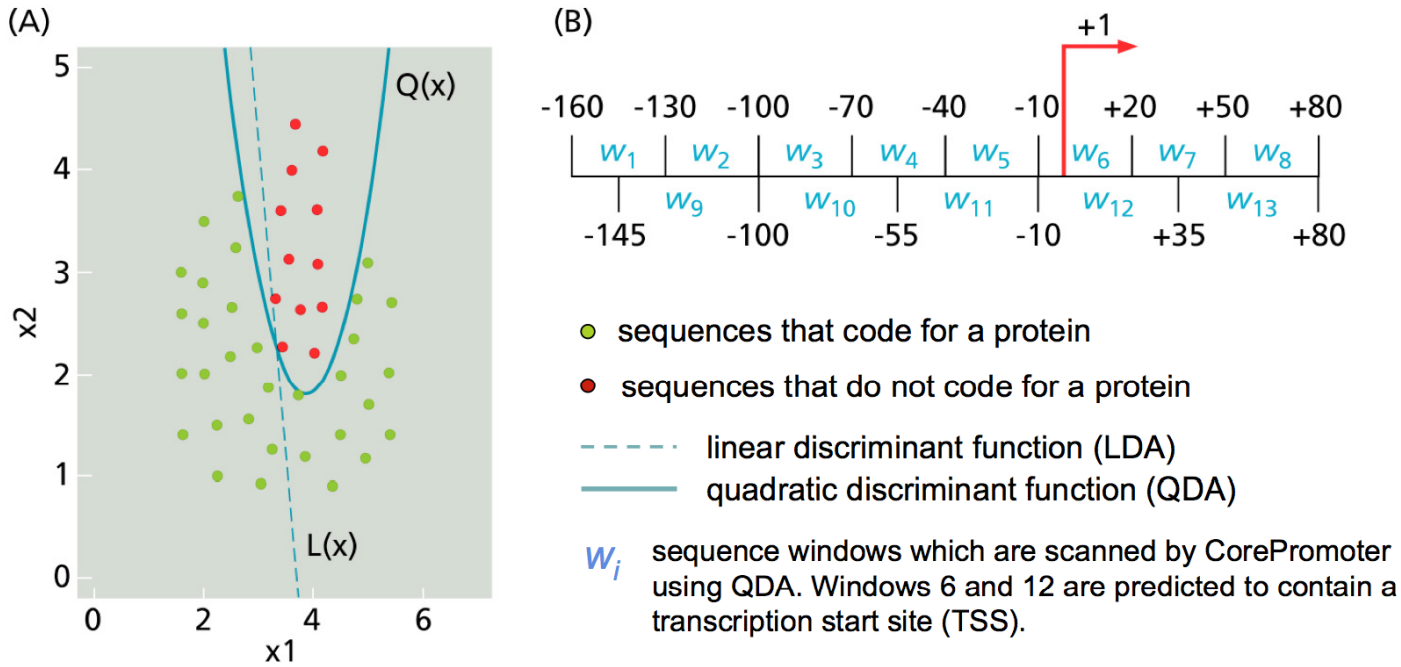


The gene finding problem

MZEF

It uses a **quadratic discriminant analysis**.

- The score are plotted for known exons and introns.
- If the score of an unknown sequence falls within the exon region then is predicted to be an exon otherwise an intron.



The gene finding problem

Evaluating predictions

Evaluation is performed exclusively on sequences where the gene structures is known. It basically relies on sensitivity and specificity measures.

- **sensitivity**: is the fraction of known genes (or bases or exons) correctly predicted.
- **specificity**: is the fraction of predicted genes (or bases or axons) that correspond to true genes.

Increasing one measure decreases the other.

The gene finding problem

Evaluating predictions

The current programs identify up to 75% of exons correctly and less than 50% of predicted gene structures correspond to actual genes [5, 6].

Table 1. Nucleotide and Exon Level Accuracy

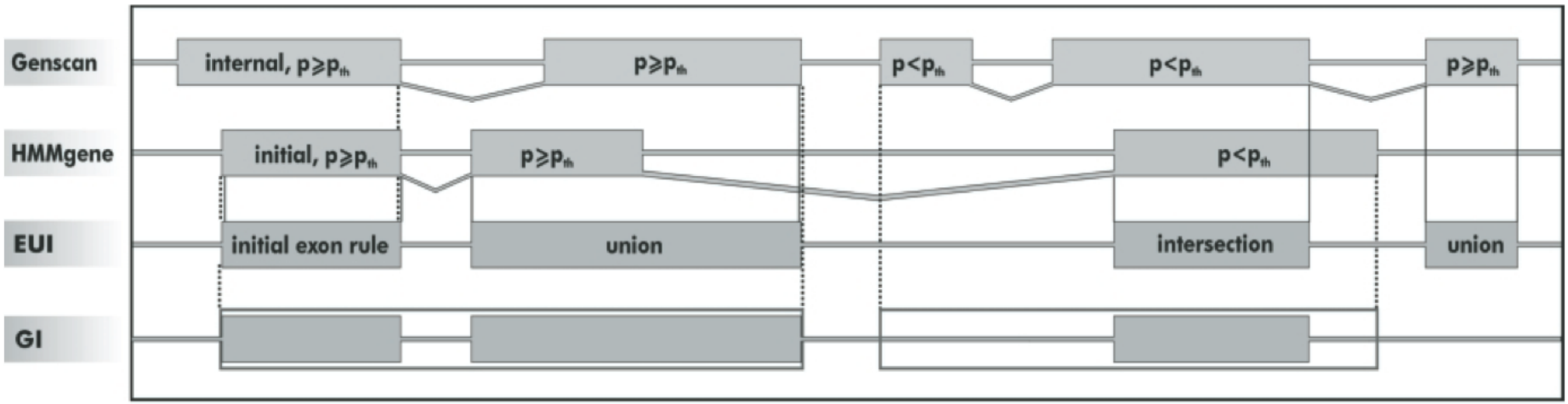
Programs	No. of sequences	Nucleotide accuracy				Exon accuracy							
		Sn	Sp	AC	CC	ESn	ESp	(ESn+ESp)/2	ME	WE	PCa	PCp	OL
FGENES	195 (5)	0.86	0.88	0.84 ± 0.19	0.83	0.67	0.67	0.67 ± 0.32	0.12	0.09	0.20	0.17	0.02
GeneMark.hmm	195 (0)	0.87	0.89	0.84 ± 0.18	0.83	0.53	0.54	0.54 ± 0.36	0.13	0.11	0.29	0.27	0.09
Genie	195 (15)	0.91	0.90	0.89 ± 0.16	0.88	0.71	0.70	0.71 ± 0.30	0.19	0.11	0.15	0.15	0.02
Genscan	195 (3)	0.95	0.90	0.91 ± 0.12	0.91	0.70	0.70	0.70 ± 0.32	0.08	0.09	0.21	0.19	0.02
HMMgene	195 (5)	0.93	0.93	0.91 ± 0.13	0.91	0.76	0.77	0.76 ± 0.30	0.12	0.07	0.14	0.14	0.02
Morgan	127 (0)	0.75	0.74	0.70 ± 0.21	0.69	0.46	0.41	0.43 ± 0.26	0.20	0.28	0.28	0.25	0.07
MZEF	119 (8)	0.70	0.73	0.68 ± 0.21	0.66	0.58	0.59	0.59 ± 0.28	0.32	0.23	0.08	0.16	0.01

Rogic et al. 2001. Evaluation of Gene-Finding Programs on Mammalian Sequences [5].

The gene finding problem

Improving predictions

Gene prediction accuracy can be improved by combining predictions from the available gene finders, a la "ensemble learning".



EUI:Exon Union Intersection, **GI**: Gene Intersection.

The gene finding problem

Improving predictions

The combined strategies improve specificity more than sensitivity at both the nucleotide and exons levels.

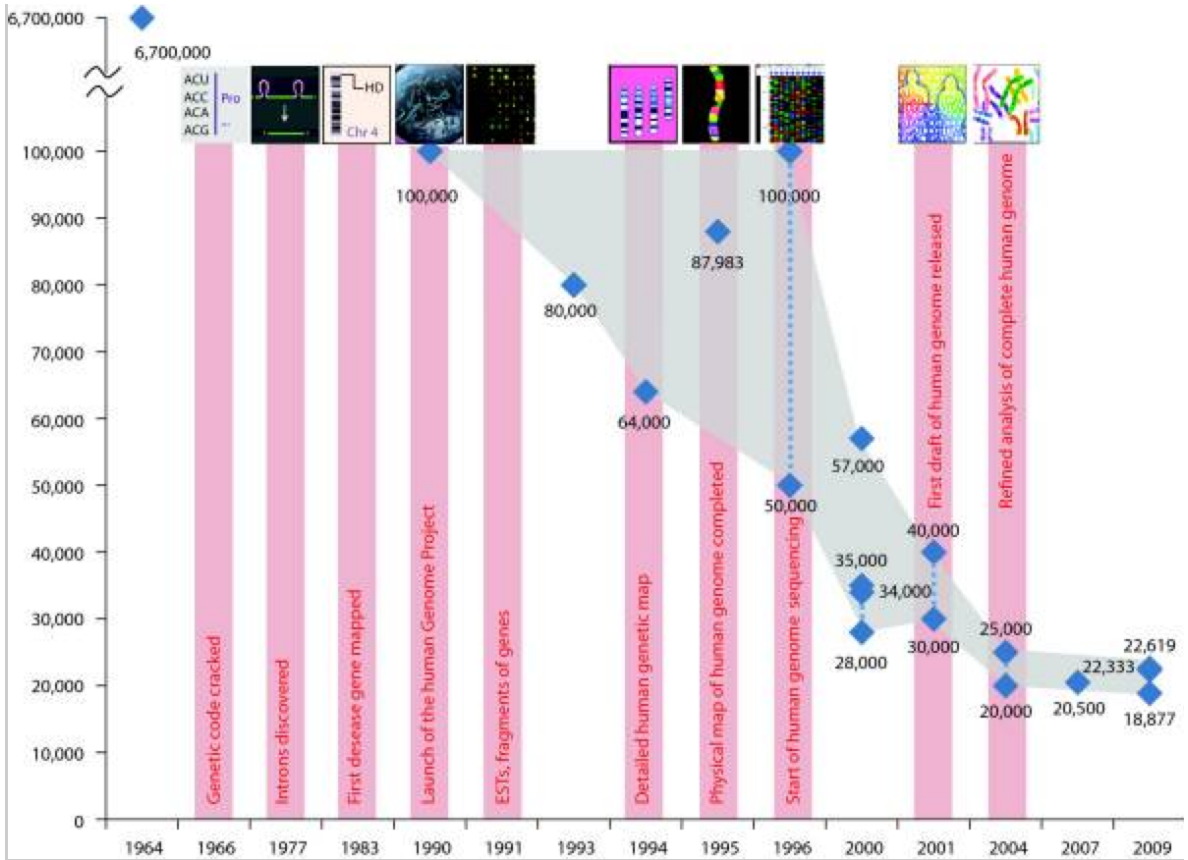
Methods	*# no prediction*	*Nucleotide accuracy*			*Exon accuracy*				
		Sn	Sp	*AC*	*ESn*	*ESp*	*(ESn + Esp)/2*	*ME*	*WE*
Genscan	3	0.95	0.90	0.91	0.70	0.70	0.70	0.08 (76)	0.09 (104)
HMMgene	5	0.93	0.93	0.91	0.76	0.77	0.76	0.12 (128)	0.07 (81)
EUI	3	0.94	**0.95**	**0.93**	**0.78**	**0.82**	**0.80**	0.10 (104)	**0.04** (55)
GI	15	0.91	**0.96**	**0.92**	**0.78**	**0.86**	**0.82**	0.19 (149)	**0.03** (43)
EUI_frame	3	0.93	**0.95**	**0.93**	**0.78**	**0.83**	**0.80**	0.11 (115)	**0.03** (46)

EUI:Exon Union Intersection, **GI**: Gene Intersection, **EUI.frame**: EUI + reading frame consistency.

How many genes in the human genome?

The trend of human gene number counts

The most successful gene finding framework for these systems was the **generalized hidden Markov model** (GHMM) approach.



Decoding the Human Genome

The ENCODE Project: ENCyclopedia Of DNA Elements

data: Since 2003, ENCODE generated more than 15 trillion bytes of raw data and consumed the equivalent of more than 300 years of computer time to analyze.

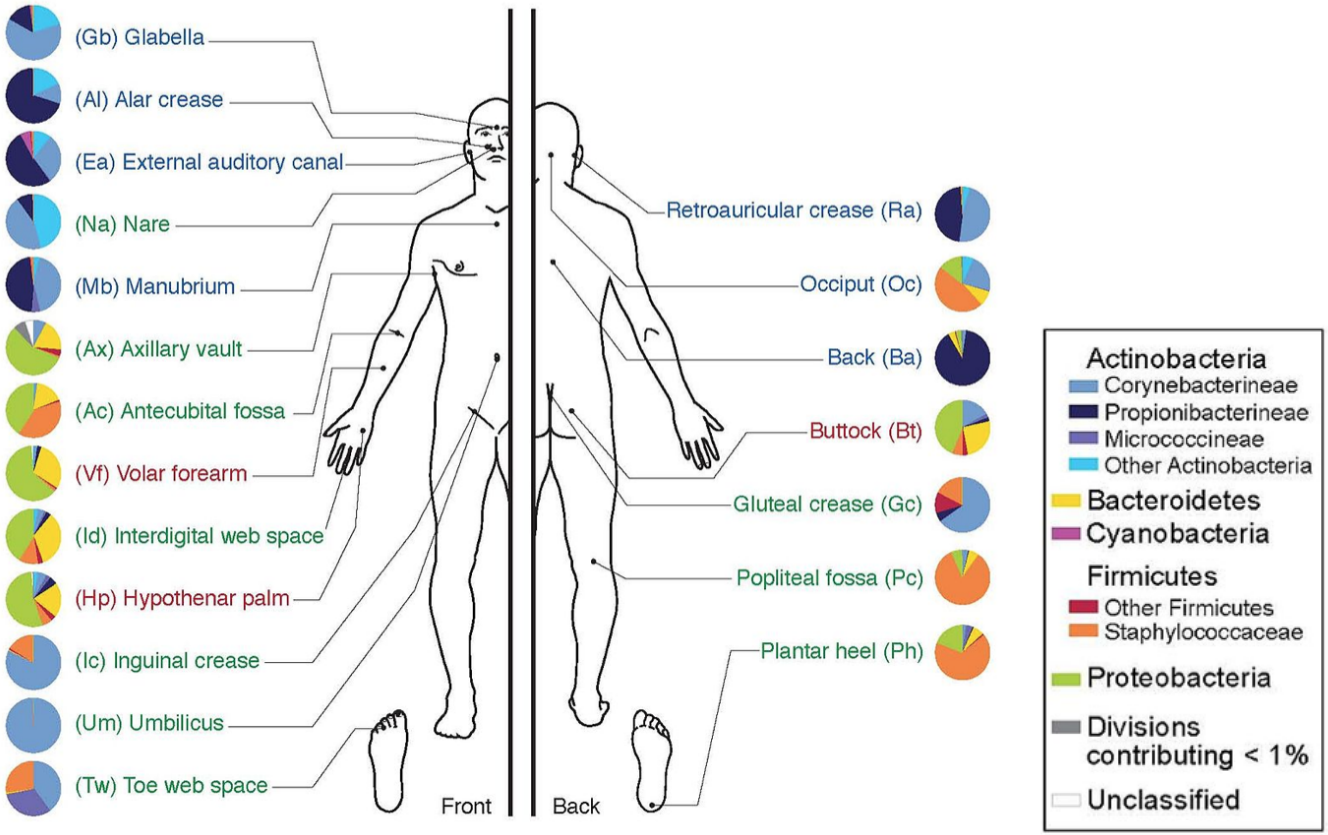
features: merge of manual and automated annotations of all human protein-coding genes, pseudogenes, and noncoding RNAs, including all splice isoforms.

`http://genome.ucsc.edu/ENCODE/`

Decoding the Human Microbiome

The Human Microbiome Project (HMP)

The total number of genes associated with the human microbiome could exceed the total number of human genes by a factor of 100-to-one.

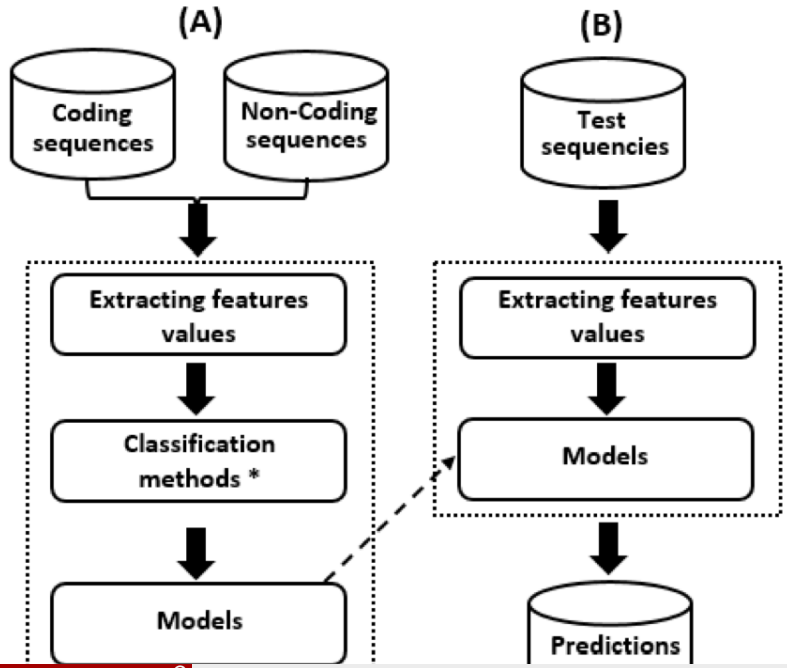


Machine Learning in metagenomics

Goes et al. (2014) Towards an Ensemble Learning Strategy for Metagenomic Gene Prediction. LNBI Springer.

i) metagenomes are a **mixture of several distinct genomes** and ii) most of the available genomes are not completed, **so mainly draft genomes** are available

ab-initio gene finders are built in a similar way as in genomics.



Machine Learning in metagenomics

A good challenge for gene finders

Feature engineering is at the core of classification strategies and it is a crucial step on prediction modeling.

	GC Content	Length	Codon usage	Dicodon usage	TIS	Aminoacid usage
Orphelia	x	x	x	x	x	
MetaGUN		x	x		x	
MGC	x	x	x	x	x	x
MetaGene	x		x	x		
FragGeneScan			x			

Content sensors features used [x] by gene prediction tools in metagenomics.

Machine Learning in metagenomics

A good challenge for gene finders

A comparison of classification methods for gene prediction in metagenomics (**Acid Mine Drainage** AMD).

Species	GenBank Acc.
Thermoplasma acidophilum *	NC_002578
Thermoplasma volcanium *	NC_002689
Acidimicrobium ferrooxidans	NC_013124
Acidithiobacillus caldus	NC_015850
Acidithiobacillus ferrooxidans	NC_011206
Acidithiobacillus ferrivorans	NC_015942
Candidatus Nitrospira defluvii	NC_014355
Thermodesulfovibrio orangestonii	NC_011296

training data: Organisms belong to the same branch of the evolutionary tree and they are associated to AMD biofilms.

Machine Learning in metagenomics

A good challenge for gene finders

A comparison of classification methods for gene prediction in metagenomics (**Acid Mine Drainage** AMD).

Species	GenBank Acc.
FA: Ferroplasma acidarmanus *	NC_021592
TA: Thermoplasmatales archaeon BRNA *	NC_020892
LFI: Leptospirillum ferriphilum	NC_018649
LFO: Leptospirillum ferrooxidans	NC_017094
SA: Sulfobacillus acidophilus	NC_015757

test data: Prokaryotic genomes associated to the same species found in AMD biolfilms.

Machine Learning in metagenomics

A good challenge for gene finders

A comparison of classification methods for gene prediction in metagenomics (**Acid Mine Drainage** AMD).

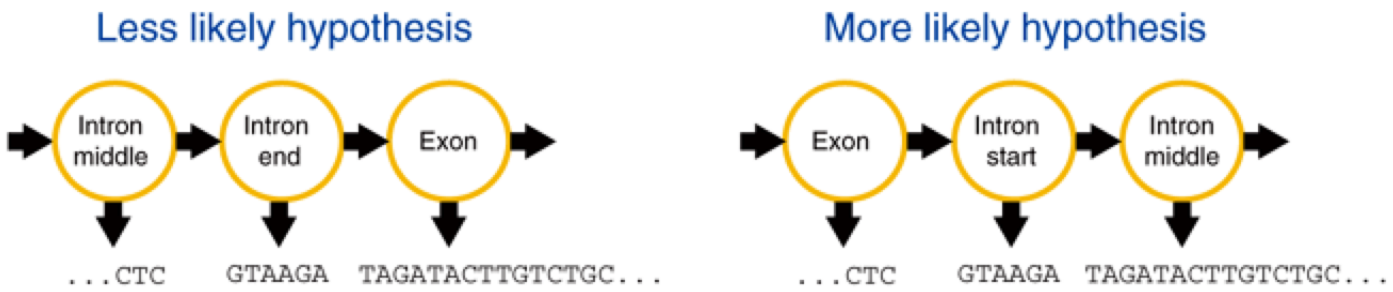
	ACC				Kappa			
Species	RF	ANN	KNN	SVML	RF	ANN	KNN	SVML
FA	0.9173	0.8702	0.8302	0.785	**0.8275**	0.7298	0.6182	0.5317
LFI	0.9156	0.9097	0.8854	0.8835	**0.8256**	**0.9097**	0.7599	0.7565
LFO	0.9263	0.9143	0.8888	0.8767	**0.8472**	0.8213	0.7666	0.741
SA	0.9383	0.9235	0.8913	0.8947	**0.8741**	0.8434	0.7746	0.7834
TA	0.957	0.9175	0.8875	0.9175	**0.9089**	0.8243	0.7577	0.737

RF:Random forest, **ANN**: Neural network, **KNN**: k-nearest neighbors, **SVML**: Support Vector Machines.

Machine Learning in (meta)genomics

Summary

Gene prediction is biased to the proper selection of potential features as well as the choice of a robust ML technique.



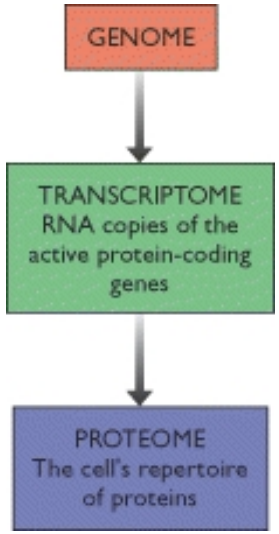
Ensemble learning techniques are getting hot in several Omics applications, though **generalized hidden Markov models** (GHMM) are in the core of several gene finders.

Machine learning in transcriptomics data

Machine learning in transcriptomics

Expression profiling

Even if every gene in a genome can be identified and assigned a function, a challenge still remains.

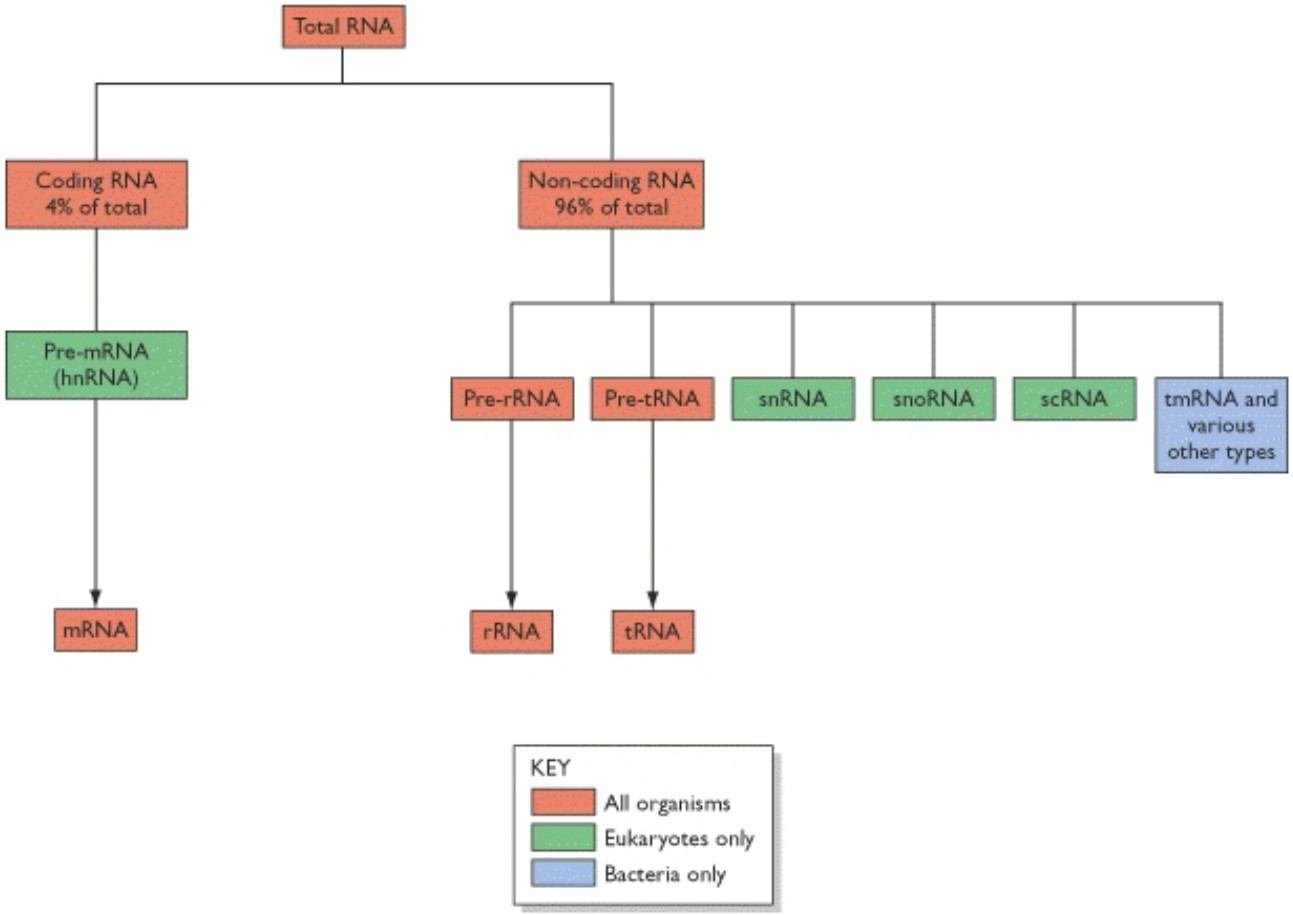


How the genome as whole operates within the cell, specifying and coordinating the various biochemicals activities that take place?

Machine learning in transcriptomics

mRNA transcripts

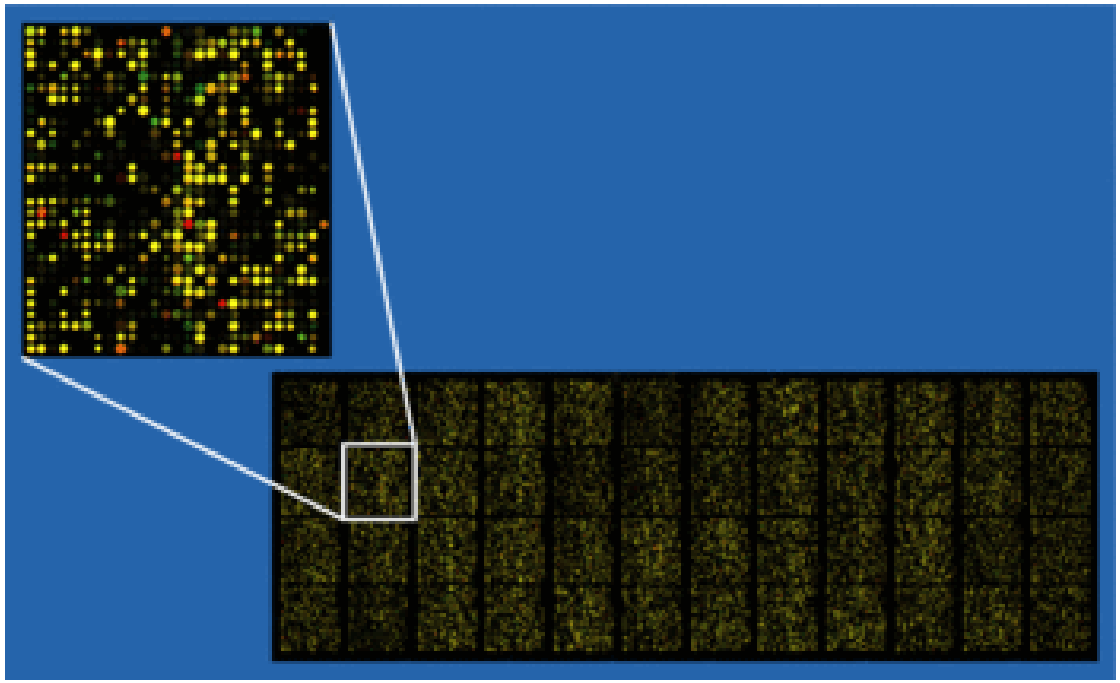
Transcriptome 4% of the total cell RNA



Machine learning in transcriptomics

Expression profiling through Microarrays technology (Closed system)

Simultaneous measure the expression of 10s of thousands of genes.

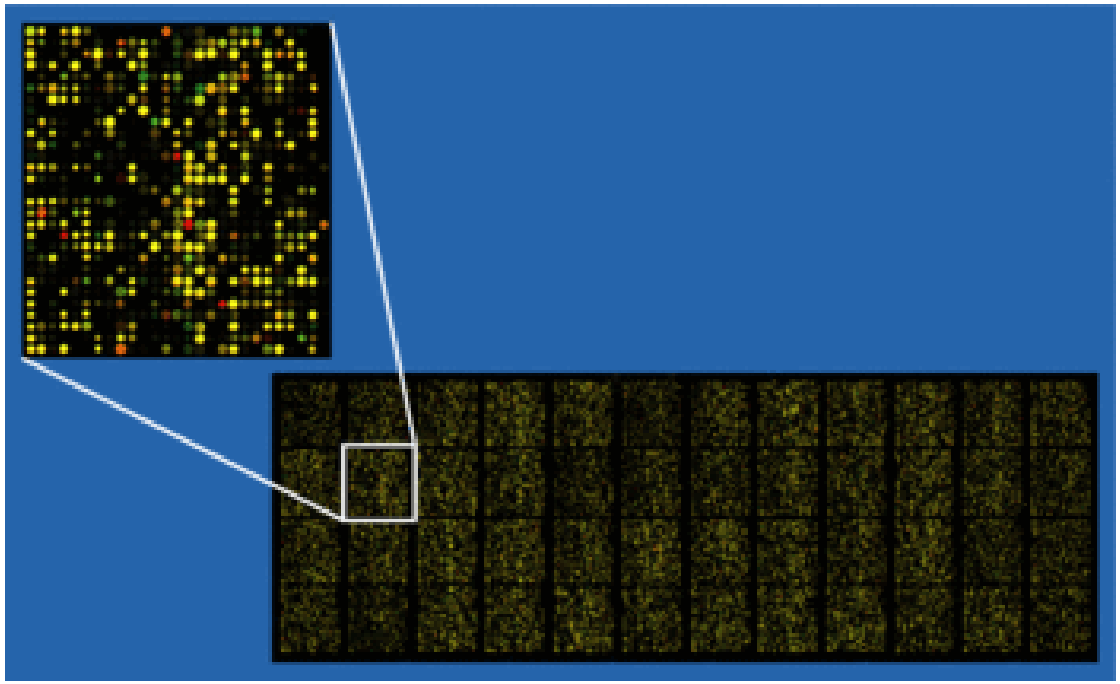


Large array of probes designed as a complementary match to the transcript of interest.

Machine learning in transcriptomics

Expression profiling through Microarrays technology (Closed system)

large p, small n dataset, where n is the number of samples and p is the number of **features** e.g. 50,000 genes, 100 patient samples is typical.



This is the opposite assumption of earlier statistical and machine learning techniques.

Machine learning in transcriptomics

Characteristics of the expression profiling data

- High dimensionality
- Sample number (n) low and observation number high (p)
- Non-independence of observations
- Complex patterns: visualisation and extraction
- Incorporation of contextual information
- Standardisation and data sharing
- Integration of with other data types

Machine learning in transcriptomics

Can lead to novel problems

- Many techniques assume $n <= p$ e.g. LDA cannot be applied directly as covariance matrix is under-determined and can not be estimated, so **feature selection** is required.
- Large opportunity for selection bias to occur in feature selection.
- Large multiple hypothesis correction problem. How to do this without being too conservative?

Even where a method e.g. SVMs can handle the high dimensionality, feature selection is still useful to remove noise genes.

Machine learning in transcriptomics

General learning tasks

- class discovery (unsupervised classification)
- class comparison (differential gene expression)
- class prediction (supervised classification)

Machine learning in transcriptomics

Class discovery

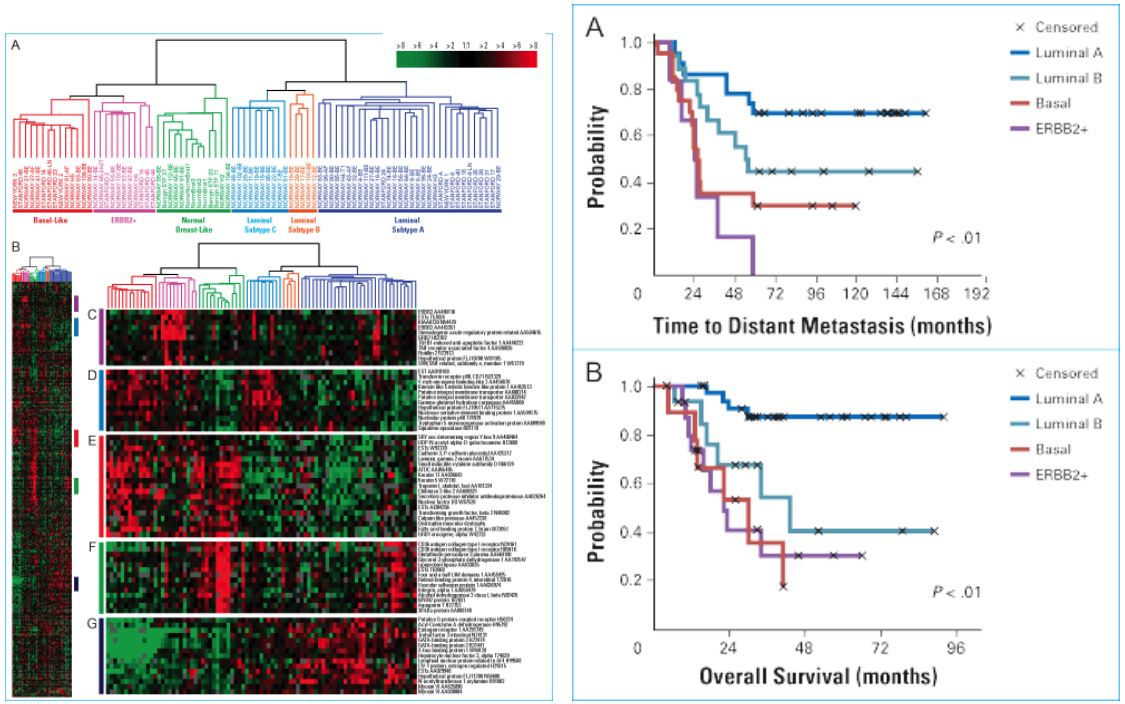
Classic Human Transcriptome Profiling Studies.

- Golub et al (1999) [3] Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. ALL acute lymphoblastic leukemia; AML acute myeloid leukemia
- Scherf et al (2000) [7] A gene expression database for the molecular pharmacology of cancer. 60 human cancer cell lines

Machine learning in transcriptomics

Class discovery

Heat maps = visualisation of gene expression profiles.



Sorlie et. al. [8] reported several previously unidentified subtypes of breast cancer using **hierarchical clustering**.

Machine learning in transcriptomics

FROM gene expression matrix TO gene distance matrix

Time	1 hr	2 hr	3 hr
g_1	10.0	8.0	10.0
g_2	10.0	0.0	9.0
g_3	4.0	8.5	3.0
g_4	9.5	0.5	8.5
g_5	4.5	8.5	2.5
g_6	10.5	9.0	12.0
g_7	5.0	8.5	11.0
g_8	2.7	8.7	2.0
g_9	9.7	2.0	9.0
g_{10}	10.2	1.0	9.2

(a) Intensity matrix, I

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

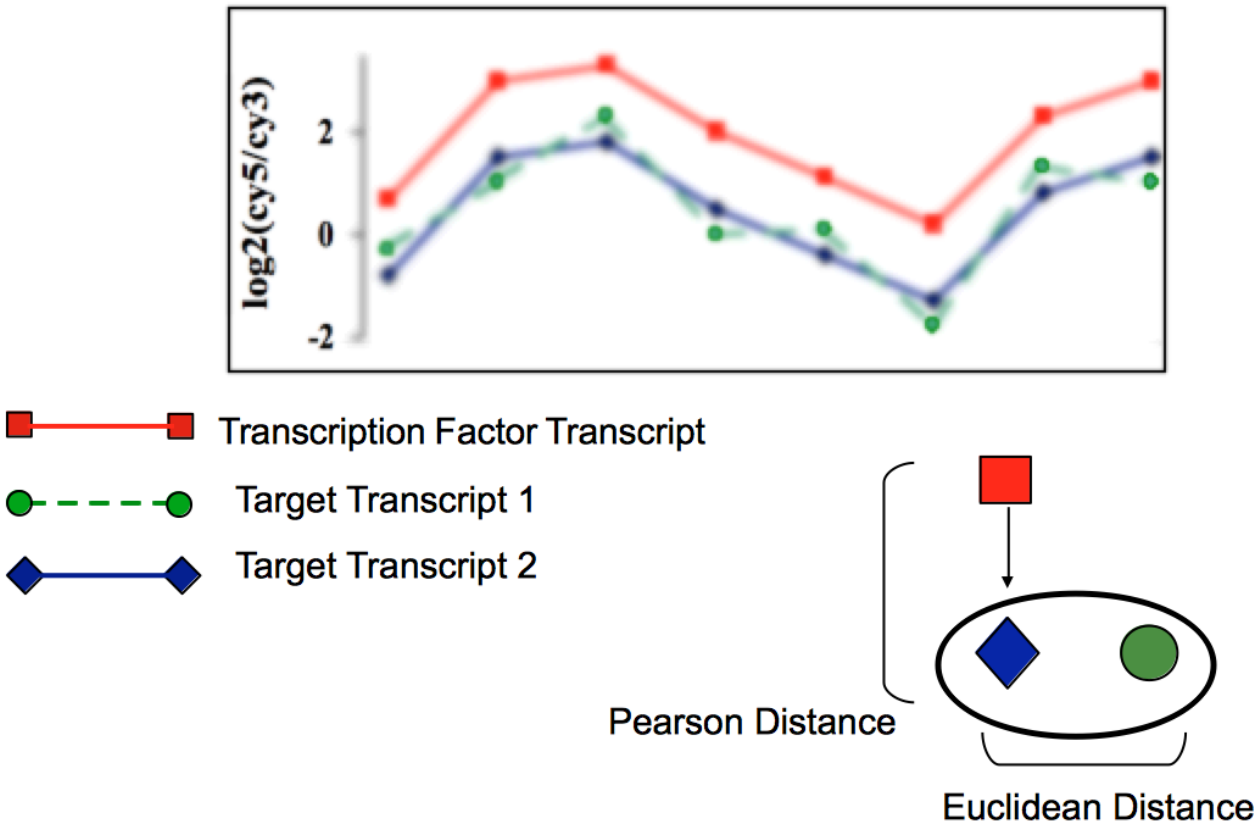
(b) Distance matrix, d

(c) Expression patterns as points in three-dimentsional space.

A cluster of genes could also be sharing functional profile, e.g. specific signalling pathway or biological annotation. **hierarchical clustering**.

Machine learning in transcriptomics

FROM gene expression matrix TO gene distance matrix

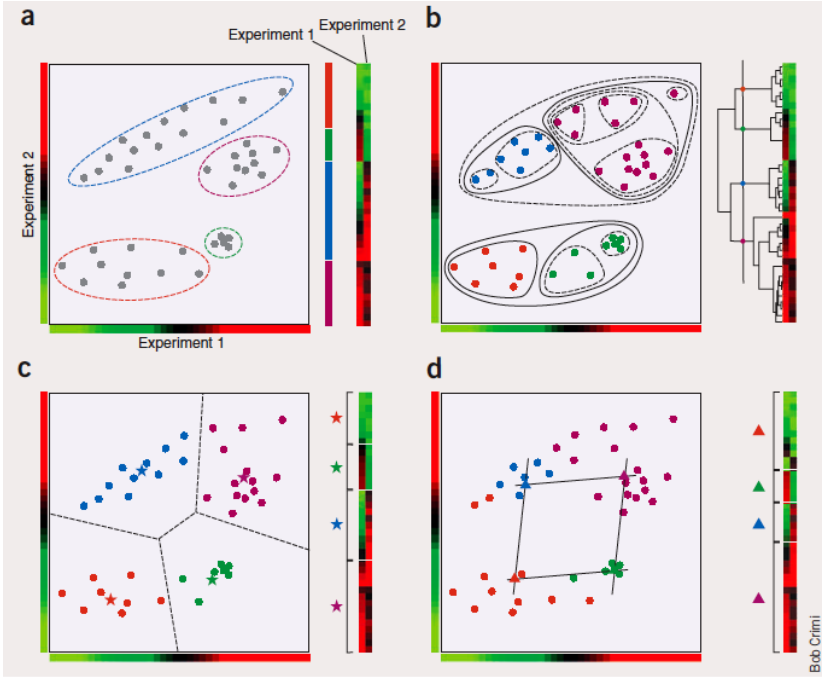


A cluster of genes could also be sharing functional profile, e.g. specific signalling pathway or biological annotation. **hierarchical clustering**.

Machine learning in transcriptomics

Class discovery

Once having the distance matrix several clustering solutions can be applied



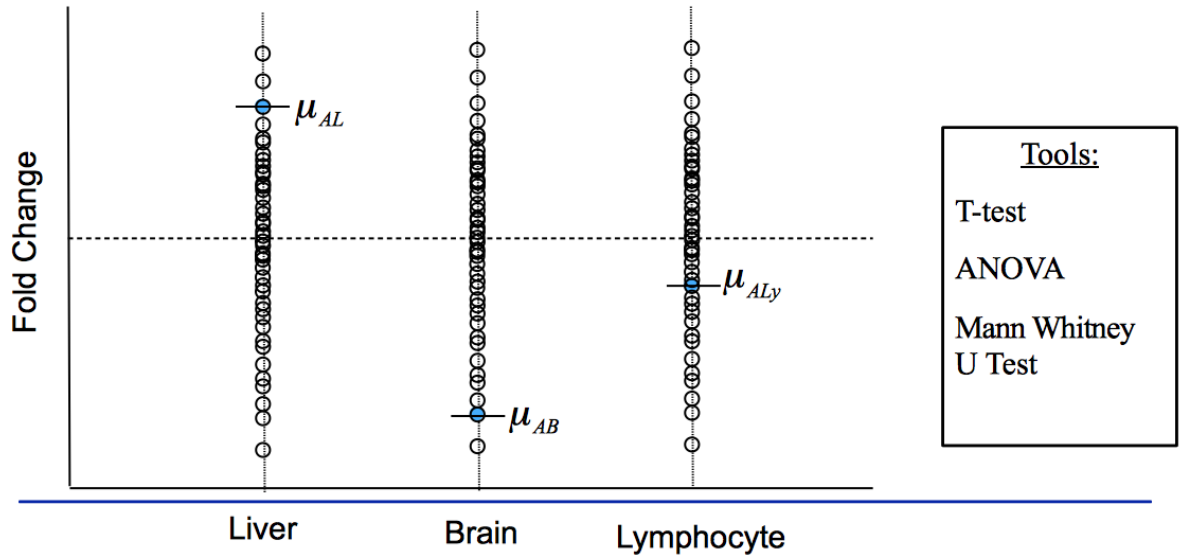
Clusters may not be associated to **biological functions**. A: Euclidean distance, B: Hierarchical clustering, C: K-means, D: SOM.

Machine learning in transcriptomics

Class comparison analysis

i) Assumes that observations are normally distributed and independent;
"Statistical significance" does not equal biological significance;
Appropriate multiple testing corrections are difficult.

$$H_0(GeneA) = \mu_{AL} = \mu_{AB} = \mu_{ALy}$$



P-value :: (P=0.01) 20,000 transcripts = 200 transcripts

Machine learning in transcriptomics

Class comparison analysis

Known also as differential expression analysis (control vs wild-type).

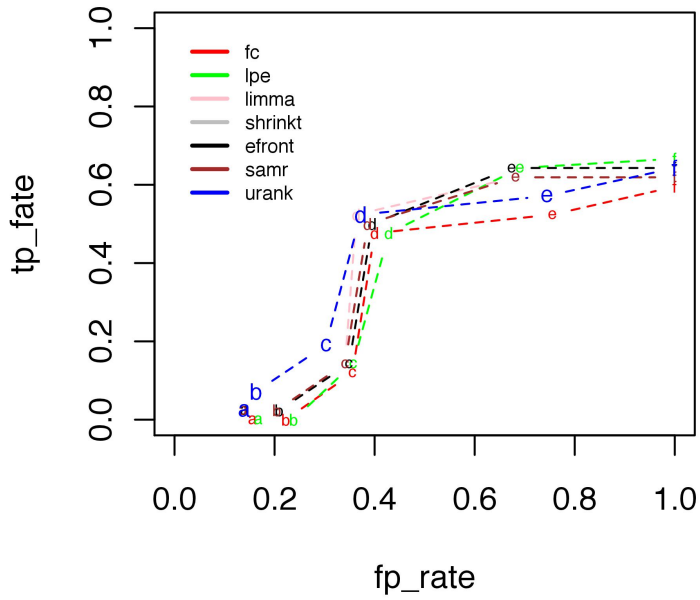
- Fold change: simplest method; ratio of expression levels (but as microarray data is typically log transformed, calculated as difference of means)
- t-statistic (one-way ANOVA F-statistic if > 2 samples) problem is that there often isnt enough data to estimate variances.
- penalised t-statistics.
- ranked gene lists : potential **gene markers**.

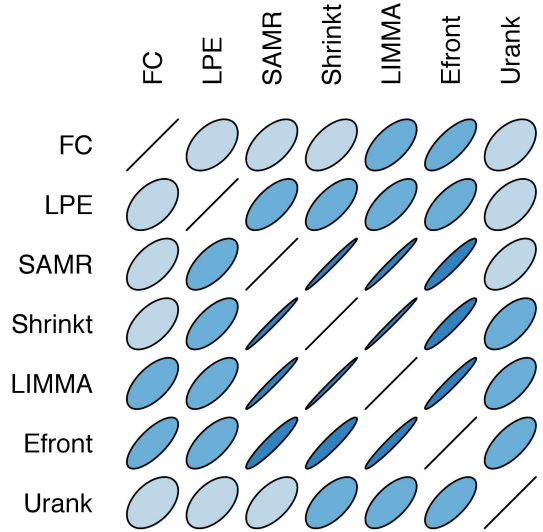
Long list of tests: LPE, Limma, Shrink-st, SAM, Fold-change, Efron-st, etc...

Machine learning in transcriptomics

Class comparison analysis

Levels of agreements between statistical tests.



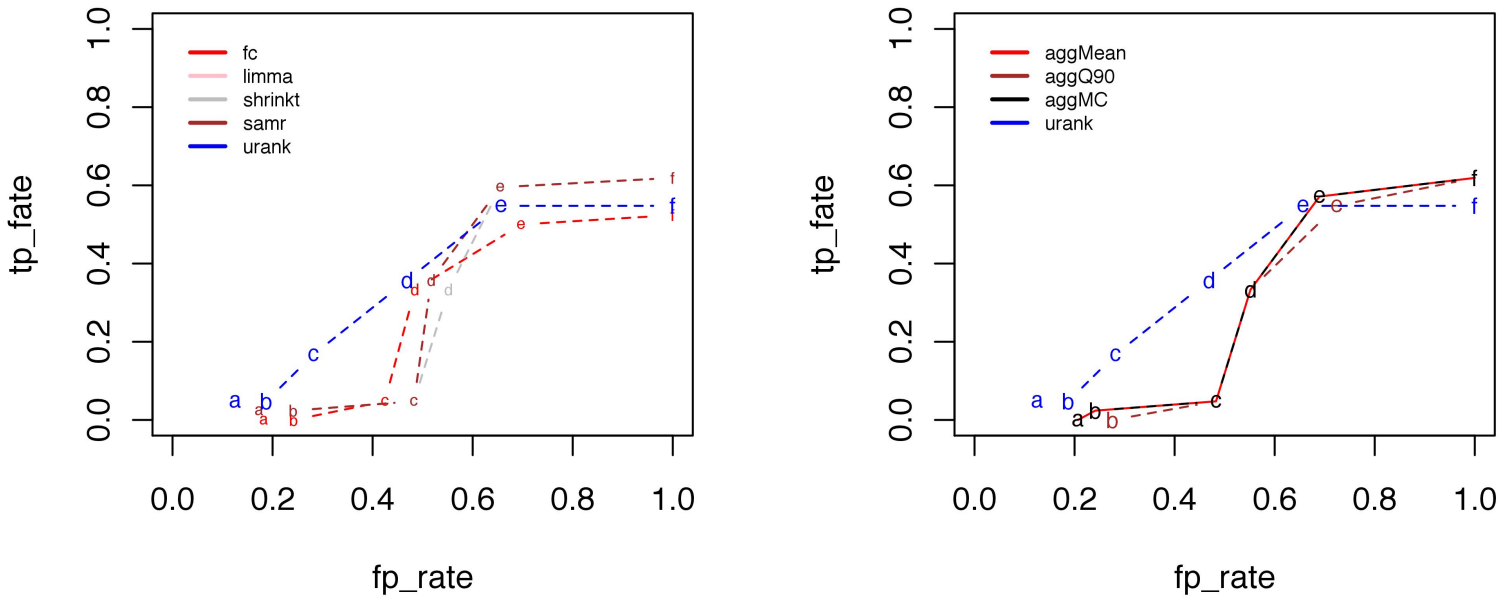


An example using the AFFY-HGU133 Affymetrix dataset.

Machine learning in transcriptomics

Class comparison analysis

Utilization of "aggregation" methods to combine gene ranks.



ROC curves of aggregation methods using the AFFY-HGU133 Affymetrix dataset.

Machine learning in transcriptomics

Class prediction

A classification problem e.g. cancer vs normal or a regression problem, e.g. survival time.

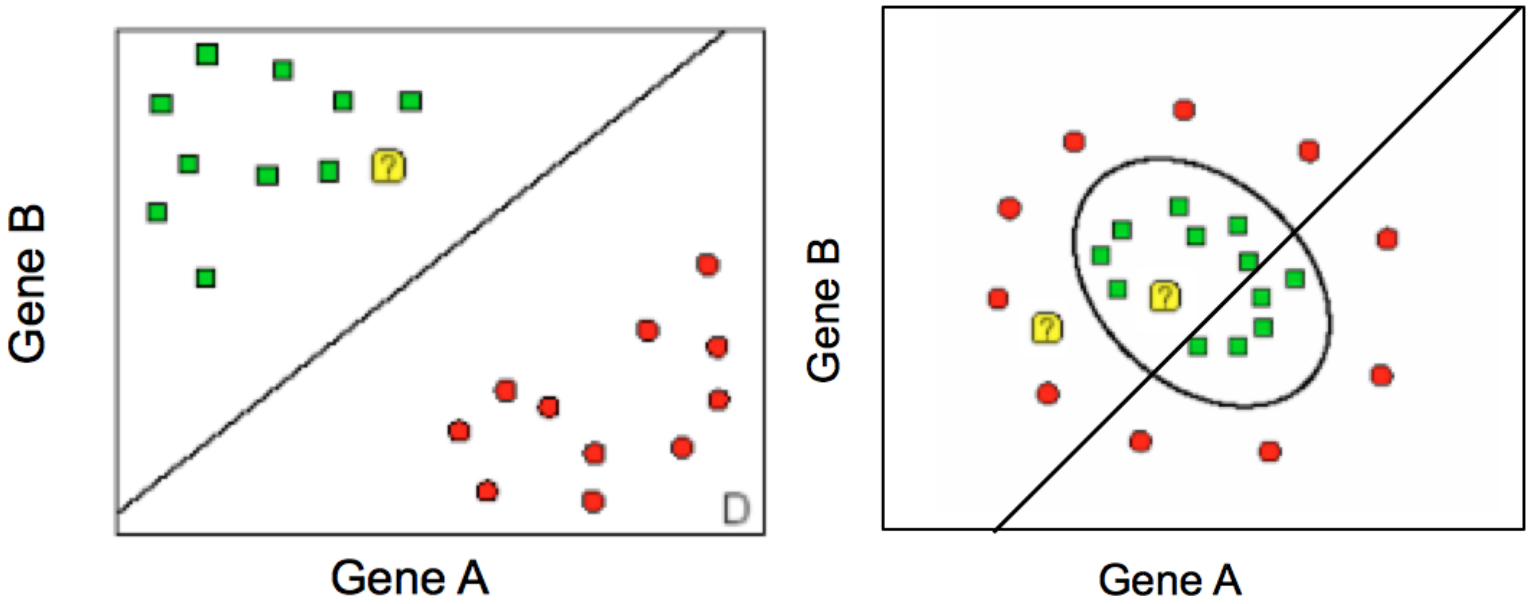
- Simple classification methods work well in practice due to small patient numbers.
- **k-nn and DLDA** performed best, and ignoring correlation between genes helped: DLDA vs correlated LDA.

Dudoit et al (2002) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data.

Machine learning in transcriptomics

Class prediction

Soft and hard decision boundaries.

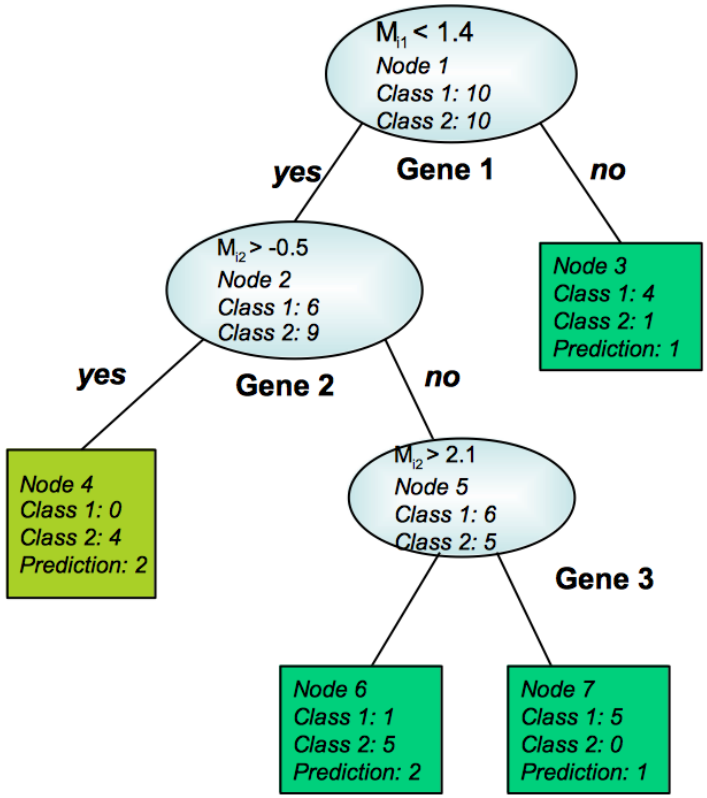


Left: linear classifiers can be good (LDA, k-nn); **Right**: more complex boundaries for linear classifiers.

Machine learning in transcriptomics

Class prediction

Rule-based exploration of features through **decision trees**.

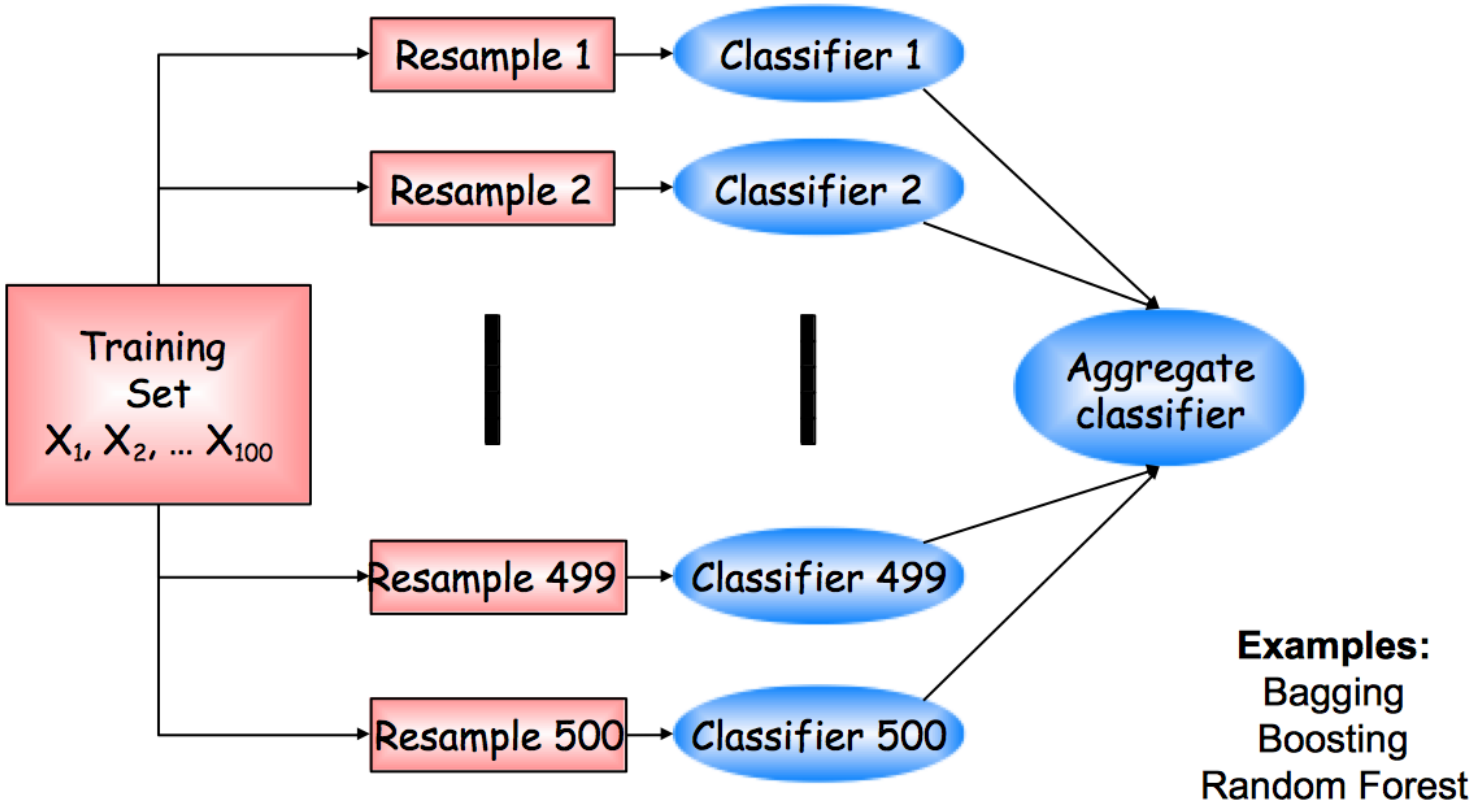


They are not sensible to hard decision boundaries.

Machine learning in transcriptomics

Class prediction

Rule-based exploration through the **aggregation of classifiers**.

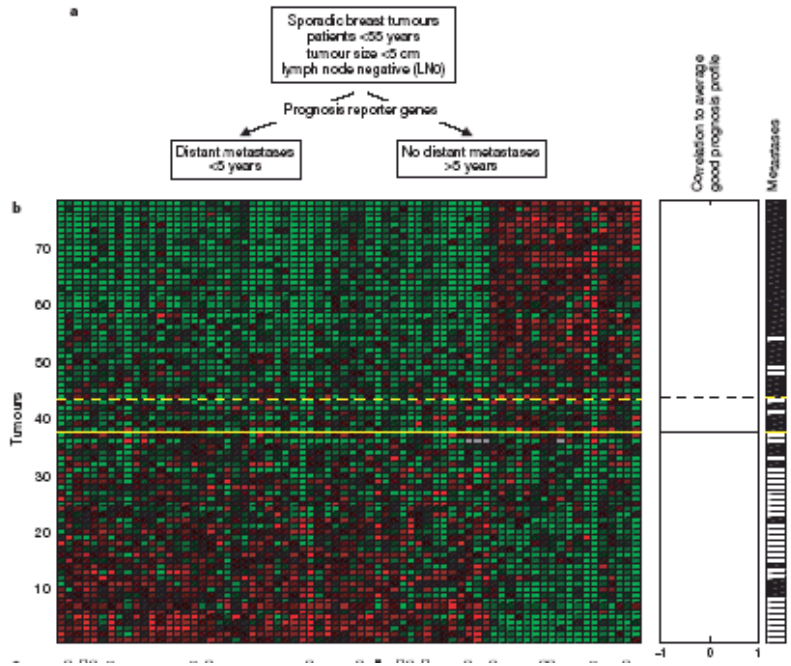


Increasing the prediction accuracy.

Machine learning in transcriptomics

Class prediction

Is the tumor ability for metastasis obtained later in development or inherent in the initial gene expression signature?

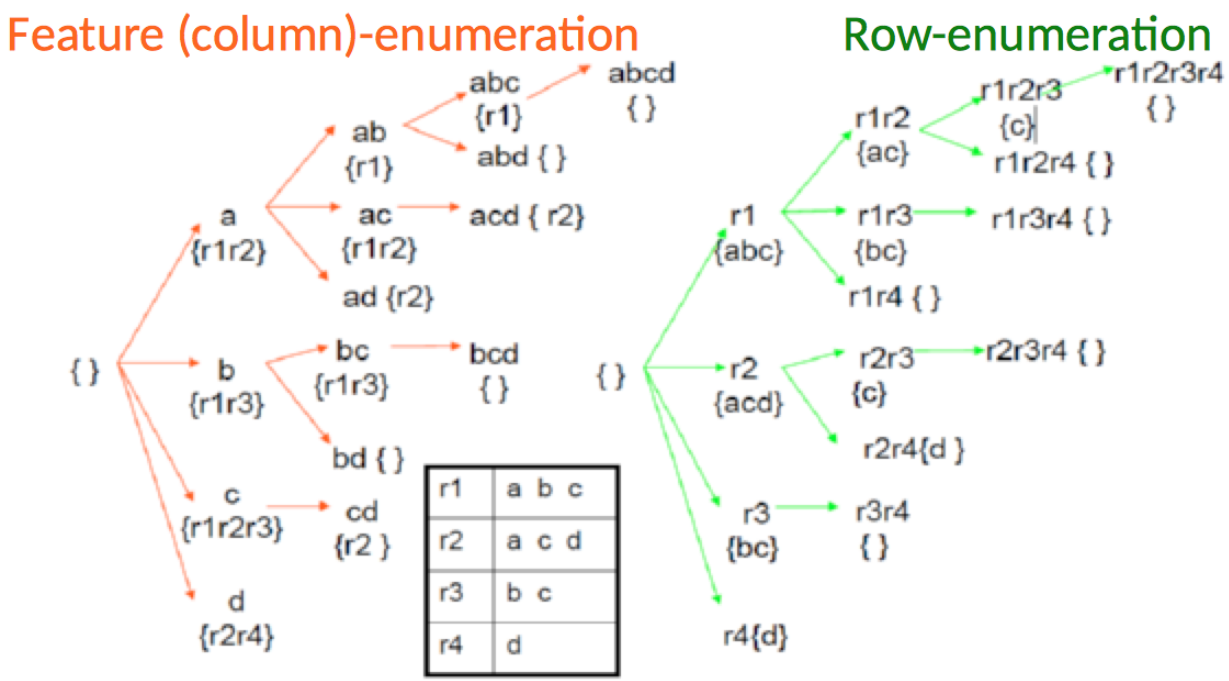


L vant Veer et al (2002) [9]. Gene expression profiling predicts clinical outcome of breast cancer. Nature.

Machine learning in transcriptomics

Class prediction

Gene associations through the enumeration of frequent patterns [1].



New methods were proposed for enumerating frequent itemsets by considering the row-space (experiments) rather than the column-space (genes).

Machine learning in transcriptomics

Class prediction

Gene associations through the enumeration of **frequent patterns**.

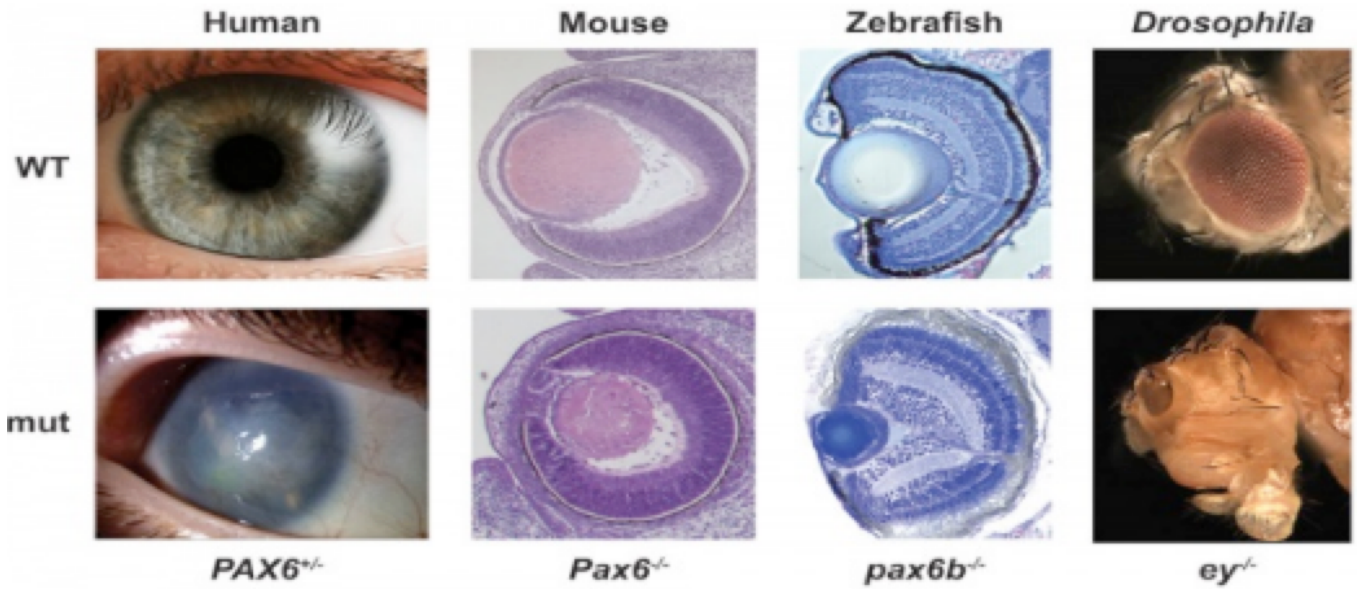
Year	Reference	Rule	Rule description
2002	Becquet *et al.* [3]	Ribosomal 150 → Cytochrome 255	When gene encoding the *ribosomal protein S24* (identified by tag 150) is overexpressed, then gene encoding the *cytochrome c oxidase subunit IV* (identified by tag 255) is also overexpressed.
2003	Creighton *et al.* [5]	NITI → ATRI, BNAI, ...	When the gene NITI is overexpressed, then a group of genes are overexpressed as well.
2005	Georgii *et al.* [40]	−STE3 › 1.2 → −SAGI › 1.1	Gene ST3 is underexpressed whenever SAGI is underexpressed as well.
2006	Carmona *et al.* [4]	Ribosome → [-]T6, [-]T7	Genes involved in the metabolic pathway *Ribosome* are underexpressed in time points 6 and 7.
2007	McIntosh *et al.* [35]	$\overline{ESC8} \rightarrow \overline{IMD1}, \overline{IMD2}$	When gene ESC8 is underexpressed then genes IMDI and IMD2 are underexpressed as well.
2008	Lopez *et al.* [7]	protein abundance = HIGH → G + C = HIGH	When the protein abundance is high, then the proportion of guanine plus cytosine in genes is high too.
2009	Nam *et al.* [43]	$POL30_{up}, YLR183C_{up} \rightarrow$ (14 minutes) $HTA2_{up}$	The overexpression of genes POL30 and YLR183c is followed by the overexpression of HTA2 after 14 min.

Types of association rules extracted from gene expression data.

Machine learning in transcriptomics

Summary

Machine learning in transcriptomics is biased to the proper selection of potential features (genes) across samples (conditions, patients, etc...).

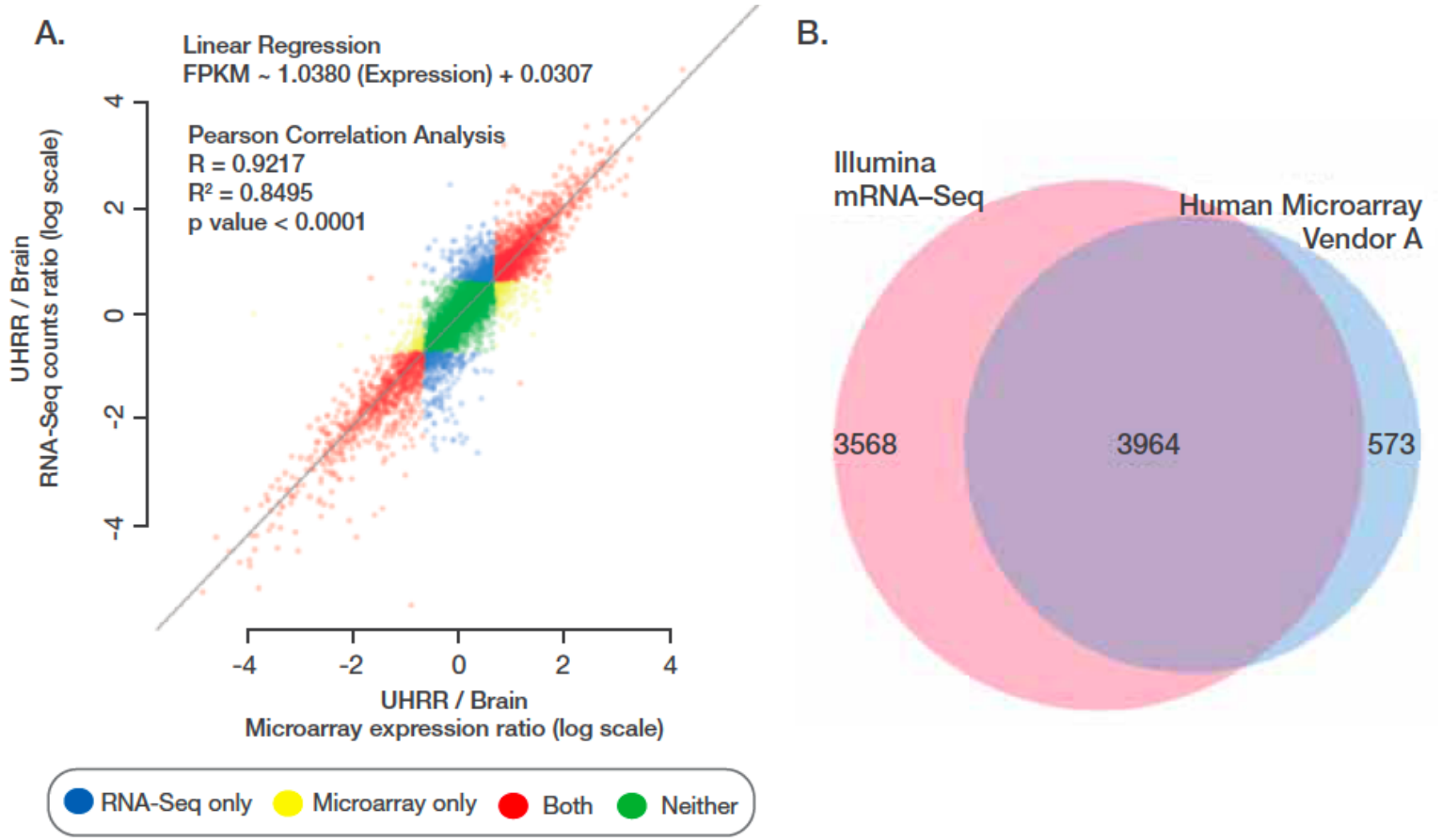


i) **Aggregation of classifiers** can deal with hard decision boundaries; ii) Statistical significance does not imply **biological soundness** of discovered patterns.

Machine learning in transcriptomics

Summary

Microarrays vs RNA-seq

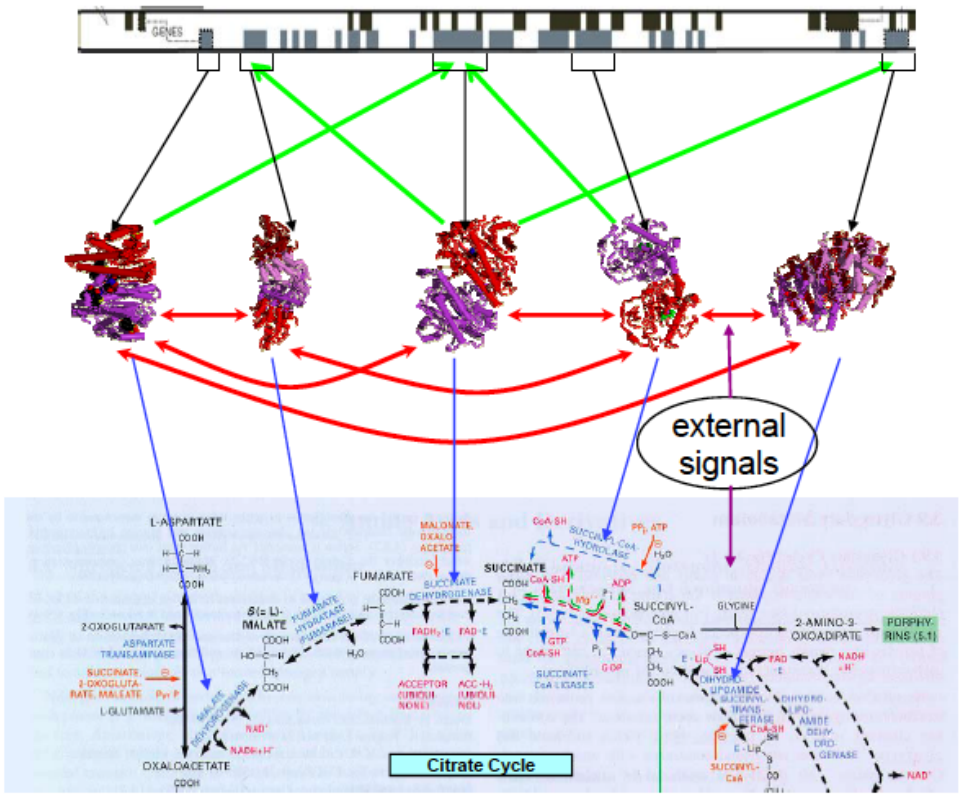


Machine learning in interactomics data

Machine learning in interactomics

Frequent biological networks

Interactomics aims to study biological networks of interactions (i.e., interactomes) between and within species in order to find how the traits of such networks are either preserved or varied.



gene regulatory network

protein-protein interaction network

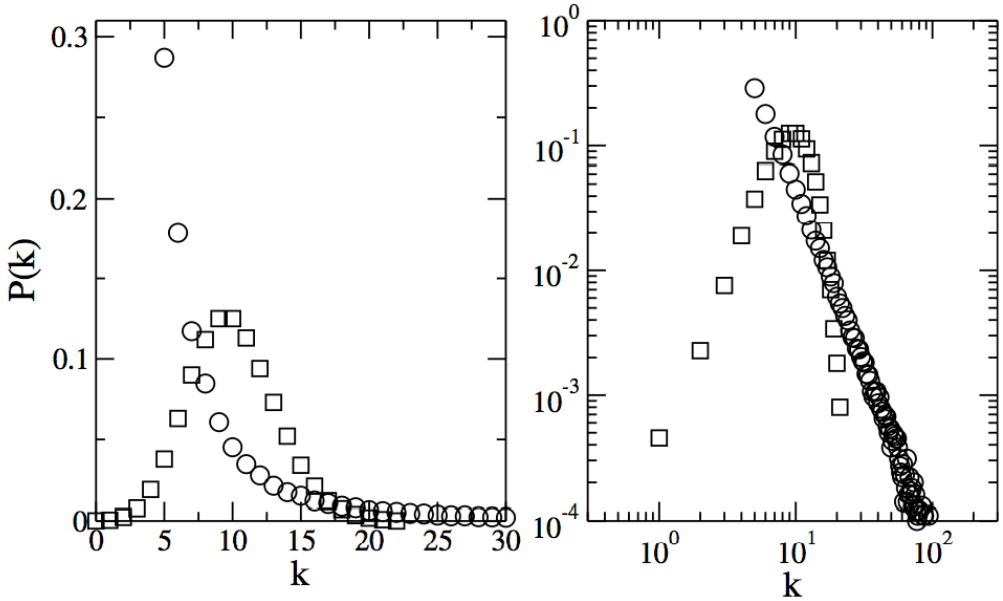
signal transduction network

bio-chemical reaction network

Machine learning in interactomics

Scale-free networks in cell biology

A cells behavior is a consequence of the complex interactions between its numerous constituents, such as DNA, RNA, proteins and small molecules.

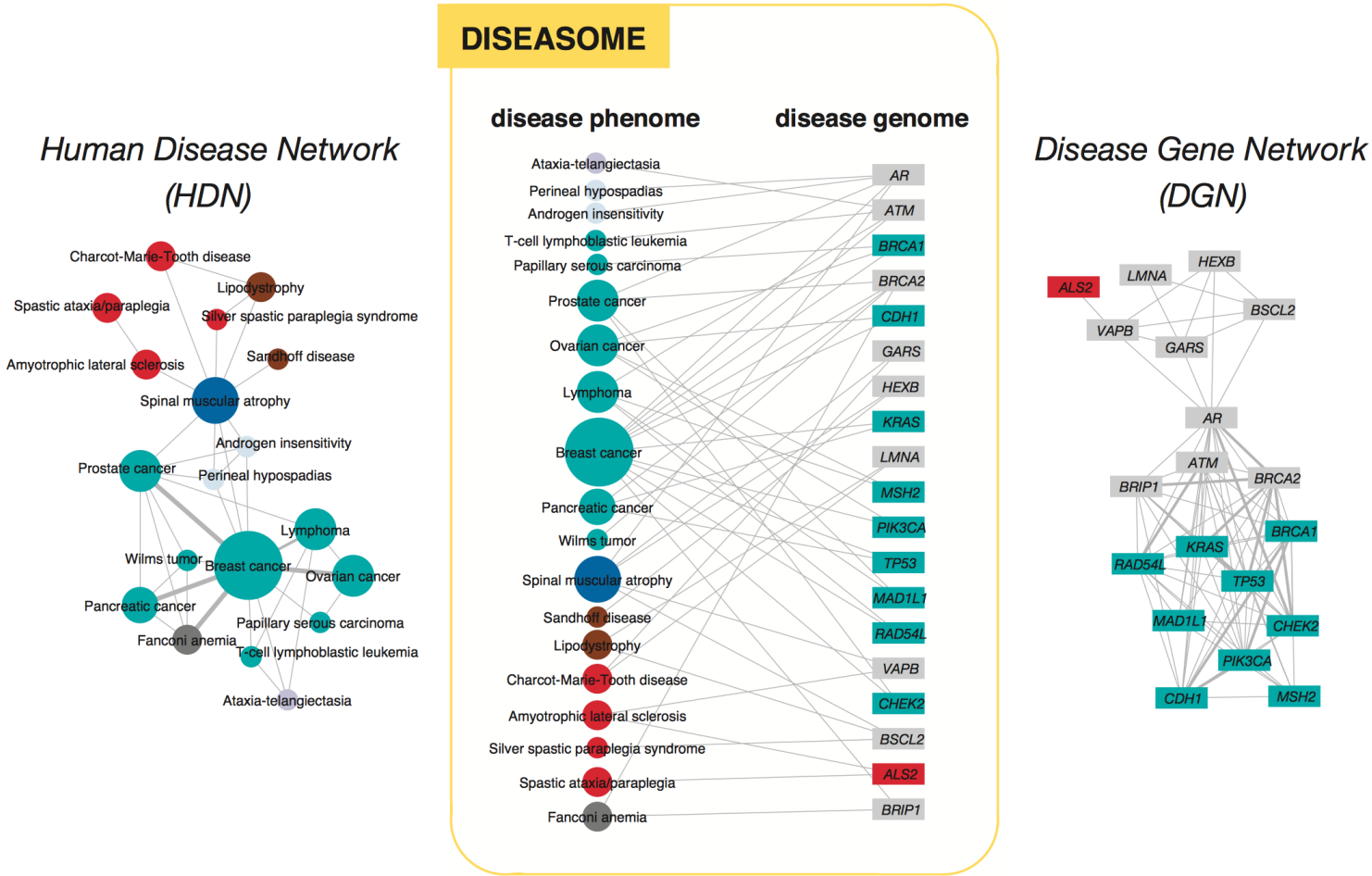


Comparison between the degree distribution of scale-free networks (○) and random graphs (□) having the same number of nodes and edges. Left (linear) and right (logarithmic).

Machine learning in interactomics

Scale-free networks in cell biology

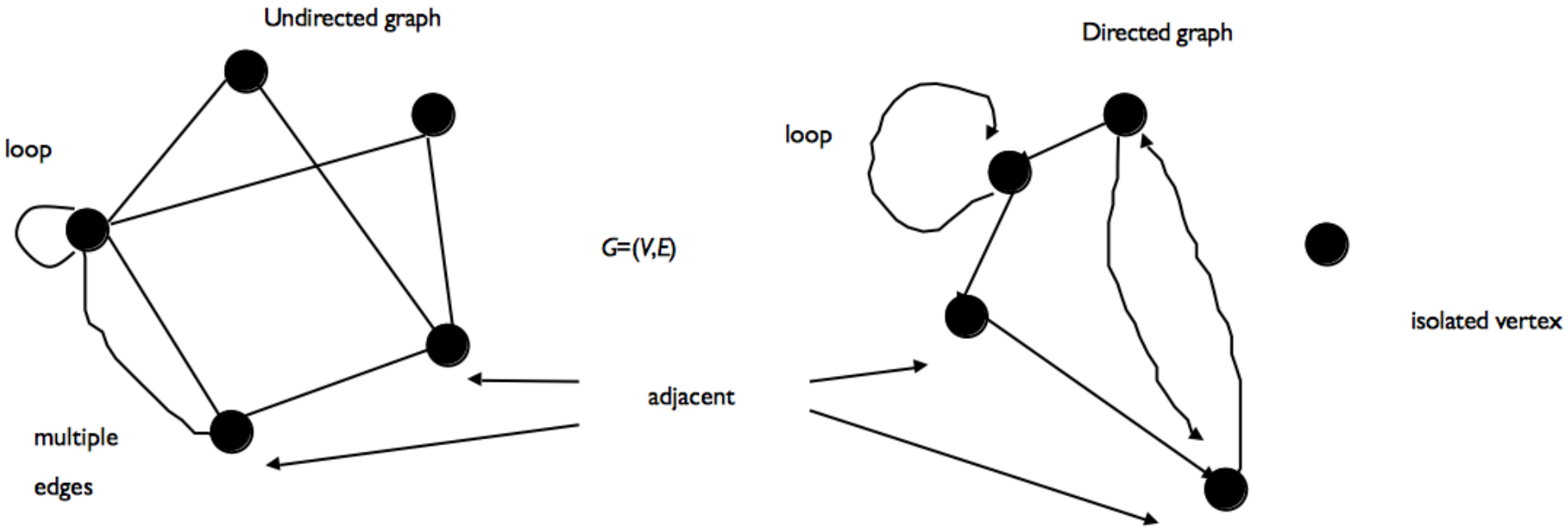
The **diseasome** bipartite network.



Machine learning in interactomics

Scale-free networks in cell biology

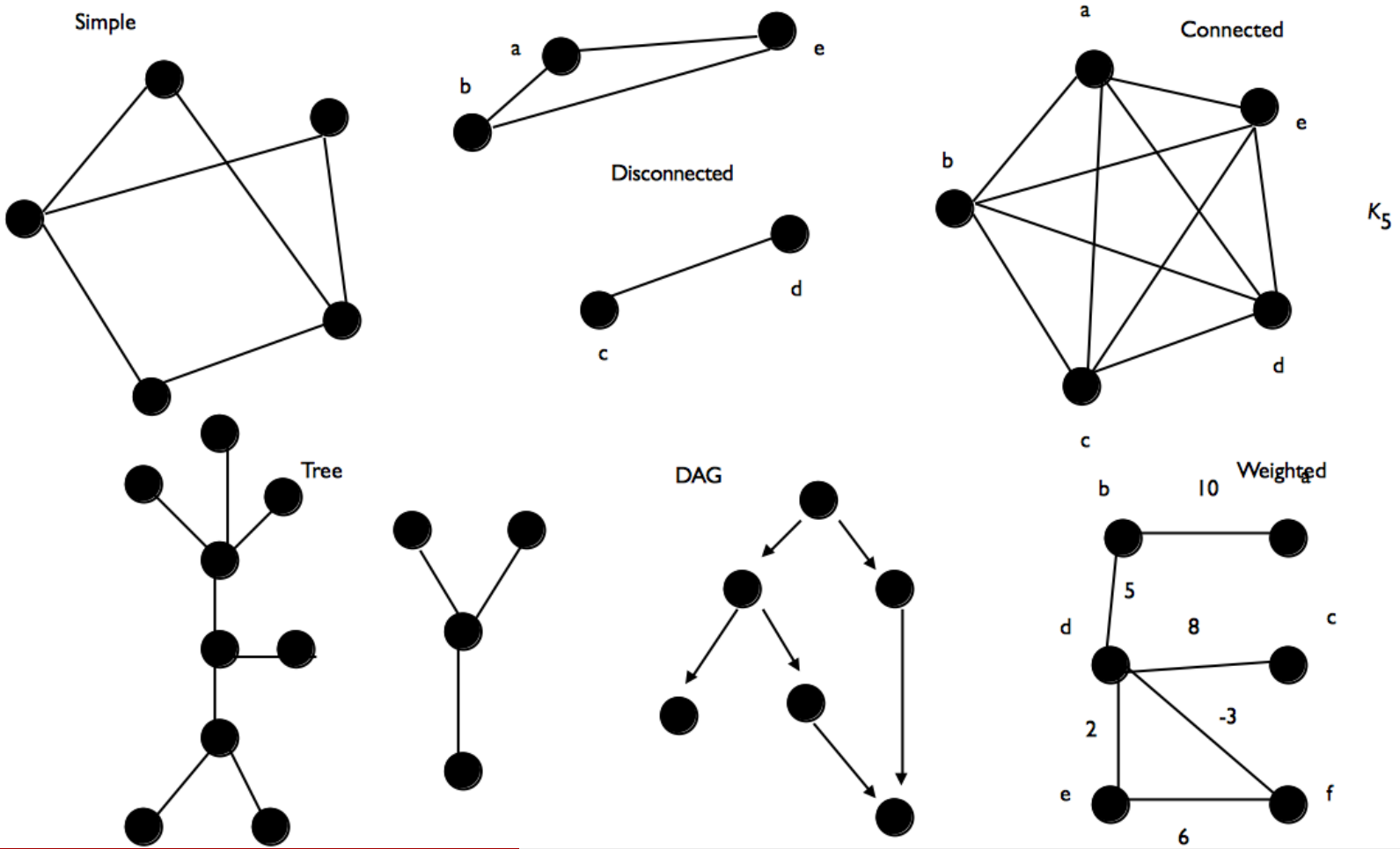
Basic graphs concepts



Machine learning in interactomics

Scale-free networks in cell biology

Basic graphs concepts



Machine learning in interactomics

High-throughput experimental methods to map interactions

- Transcription factors bind to the promoter regions of genes. They have a **DNA binding domain** and an activation domain.
- In the **two-hybrid method** the two domains are separated, and fused to two proteins. If the two proteins interact by binding, the transcription factor activates the expression of a reporter gene.
- Systematic experiments with all proteins in a given organism lead to **genome-wide protein interaction maps**.

Machine learning in interactomics

High-throughput experimental methods to map interactions

Protein networks: i) Co-affinity purification + mass spectrometry, ii) Yeast two hybrid

Databases:

- Database of Interacting Protein (DIP),
- the Biomolecular Interaction Network (BIND),
- the Munich Information Center for Protein Sequences (MIPS),
- the Human Protein Reference Database (HPRD), and the Yeast Proteome Database (YPD)

Machine learning in interactomics

High-throughput experimental methods to map interactions

Metabolic networks: i) Enzyme characterizations: protein and DNA microarrays, ii) Metabolite identification: isotope labeling, iii) Flux quantification: Mass spectroscopy.

Databases:

- Kyoto Encyclopedia of Genes and Genomes (KEGG),
- Ecocyc,
- MetaCyc

Machine learning in interactomics

High-throughput experimental methods to map interactions

Transcriptional regulatory networks: i) DNA footprinting, ii) chromatin immunoprecipitation (ChIP)

Databases:

- Kyoto Encyclopedia of Genes and Genomes (KEGG),
- Transcription Factor Database (TRANSFAC),
- Regulon Database (RegulonDB)

Machine learning in interactomics

Sizing protein interaction networks

The topological properties of diverse protein interaction networks are similar.

Studies

- Ito (yeast): 8868 interactions between 3280 proteins
- Uetz (yeast): 4480 interactions, 2115 proteins
- Giot (Drosophila): 4780 interactions among 4679 proteins
- Li (C. elegans): 5534 interactions, 3024 proteins
- Rual (human): 2800 interactions, 8300 proteins

S.-H Yook, Z.N. Oltvai, A.-L. Barabasi, Proteomics 4, 928 (2004)

Machine learning in interactomics

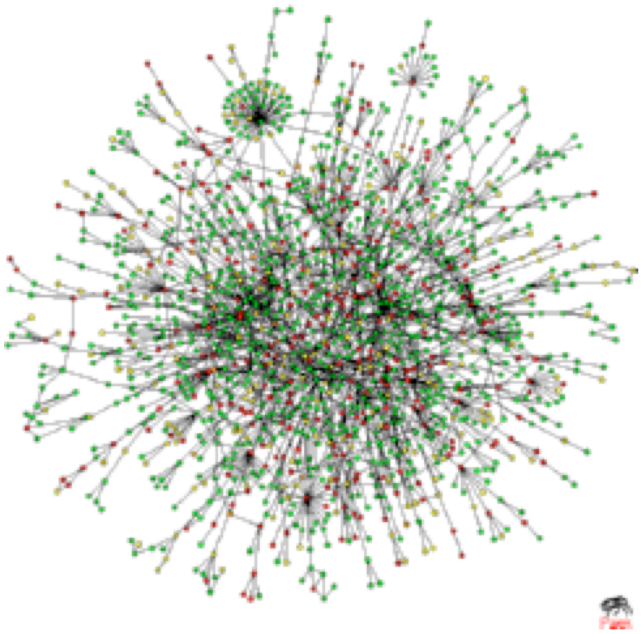
Biological network properties

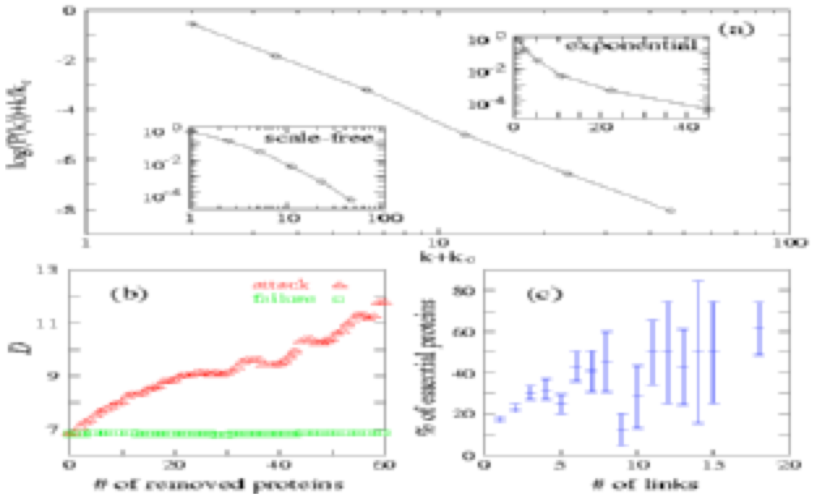
- **Power law** degree distribution: Rich get richer
- Small World, A small average path length: Mean shortest node-to-node path
- Robustness, Resilient and have strong resistance to failure: On random attacks and vulnerable to targeted attacks
- **Hierarchical Modularity**: A large clustering coefficient How many of a nodes neighbors are connected to each other.

Machine learning in interactomics

Biological network properties

Yeast protein interaction show the power law property.





$$P(k) \sim (k + k_0)^{-\gamma} \exp(-\frac{k + k_0}{k_\tau})$$

H. Jeong, S.P. Mason, A.-L. Barabasi Z.N. Oltvai, Nature, 2001

Machine learning in interactomics

Biological network properties

Clustering coefficient C is defined as the probability that two neighbors of a given node are adjacent.

$C_v = 2E_v/d_v(d_v - 1)$
where
E_v is the number of edges between neighbors of v.
A node v has d_v neighbors.

The clustering coefficient C of the whole network is the average of C_vs for all nodes v in the network.

"all-my-friends-know-each-other" property

Machine learning in interactomics

Biological network properties

Shortest path **edge betweenness**, computes the fraction of shortest paths passing through an edge.

$eb(e_ij) = SP_{ij}/SP_{max}$

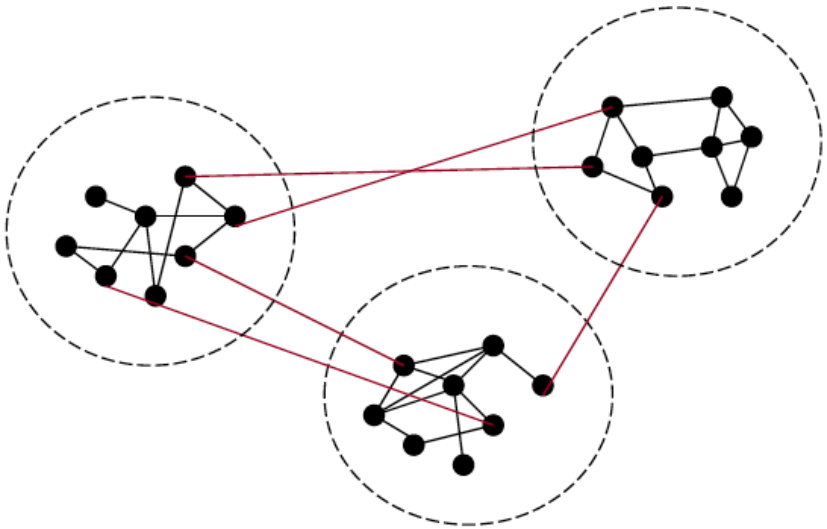
Edges that lie between communities have high values of betweenness.

"I-am-between-every-pair" property

Machine learning in interactomics

Biological network properties

Many real-world networks, especially social ones, exhibit **community structure** (also called modularity)
Modules are potentially involved in common cellular functions or protein complexes



Intuitively community structure can be defined as the existence of **subgraphs** that are densely connected but sparsely inter-connected.

From gene-expression To gene co-expression networks

Alves et al. (2013) A Network-Based Meta-analysis Strategy for the Selection of Potential Gene Modules in Type 2 Diabetes. BSB'2013.

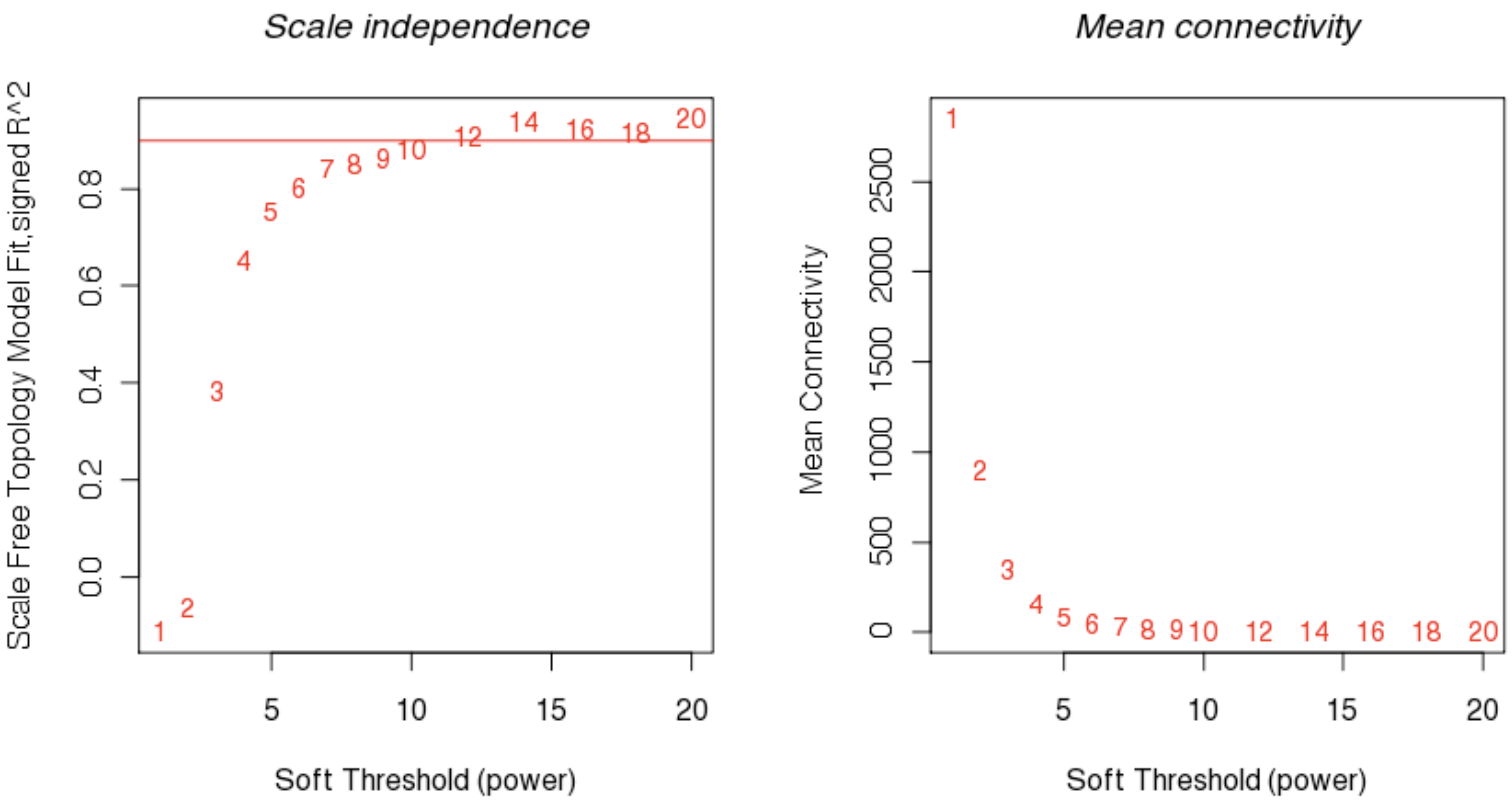
Affymetrix chips obtained in Gene Expression Ominibus (GEO).

Study	Samples	Genes before	Genes after	Affy.Chip
GSE12389	8	45101	5316	Mouse430_2
GSE2253	20	22690	11686	Mouse430A
GSE12639	12	31099	14998	Rat230_2
GSE13270	101	31099	13935	Rat230_2

Preprocessing

Robust Multi-array Average (RMA) followed by the detection of present and marginal gene calls in at least 75% of all samples

Weighted gene co-expression networks



Network topology

The soft-thresholding strategy, adopted by *WGCNA*, keeps all possible links and raises the original coexpression values to a power "beta" so that the high correlations are emphasized at the expense of low correlations

Using the TOM matrix to cluster genes

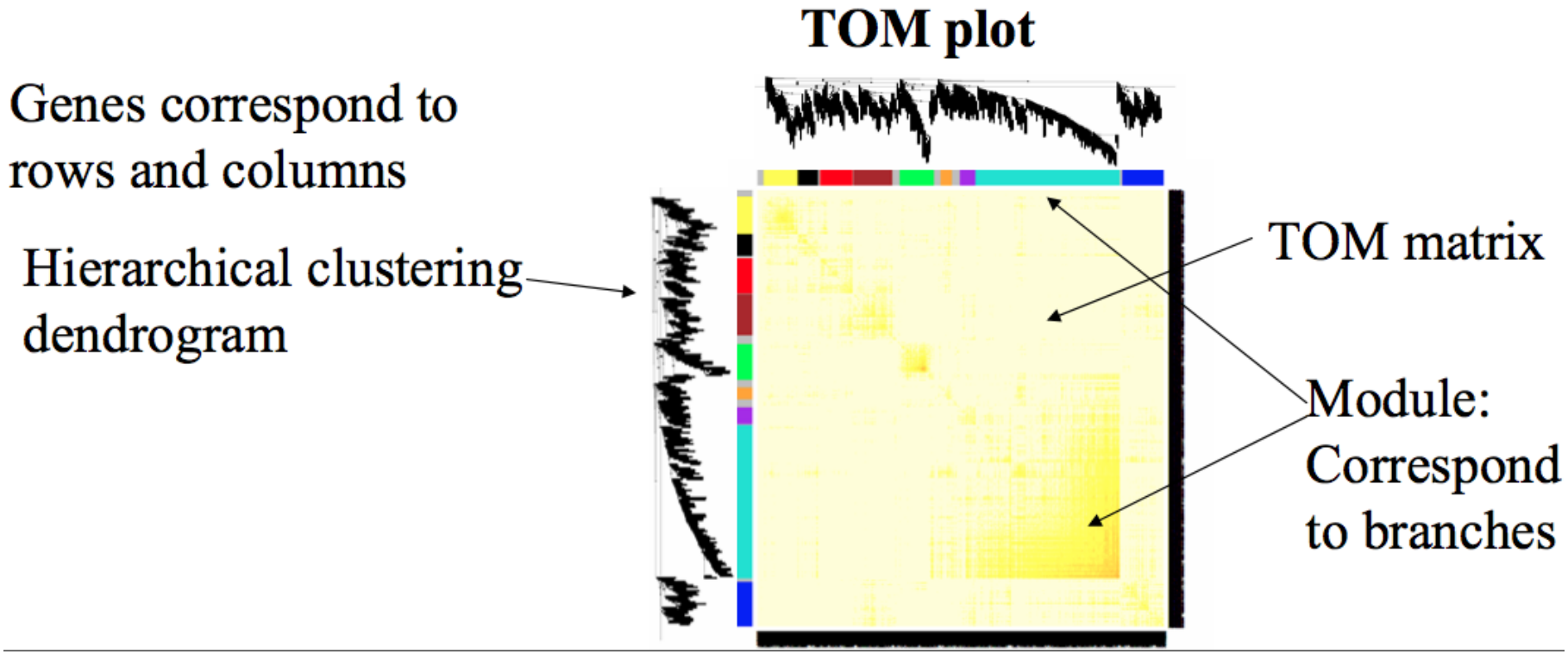
$$TOM_{ij} = \frac{\sum_{u} a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$DistTOM_{ij} = 1 - TOM_{ij}$$

Topological overlap measure

Generalized in Zhang and Horvath(2005) to the case of weighted networks

TOM leads to a network distance measure



Network modules

To group nodes with high topological overlap into modules (clusters), we typically use average linkage hierarchical clustering coupled with the TOM distance measure networks

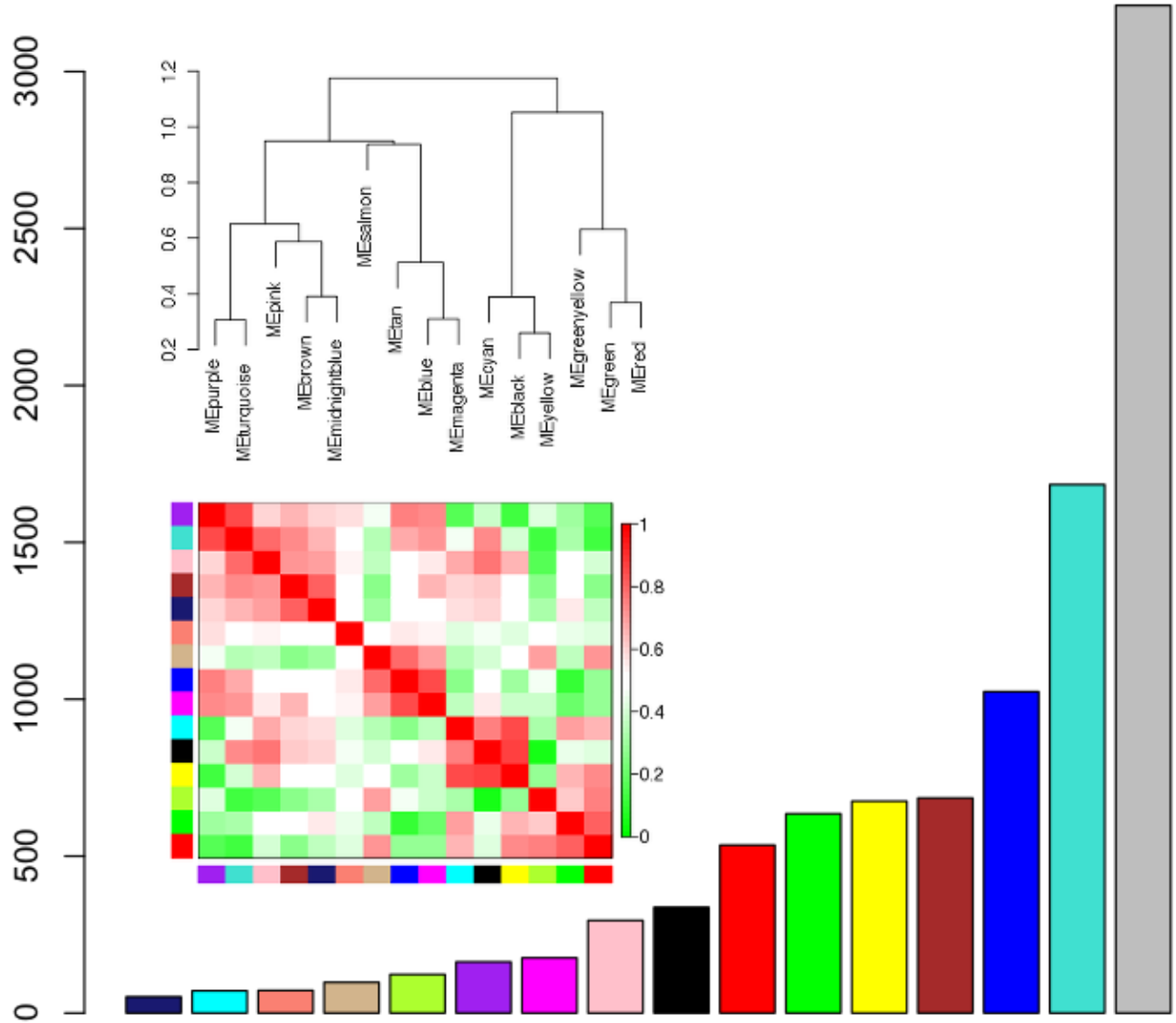
Gene modules

Study	Modules	Genes before	Genes after	Affy.Chip
GSE12389	12	5316	1566	Mouse430_2
GSE2253	17	11686	5570	Mouse430A
GSE12639	34	14998	5660	Rat230_2
GSE13270	16	13935	9836	Rat230_2

Hub genes

Intramodular hub genes are highly correlated with the module eigengene. A gene in a module is considered significant if it has a strong p-value ($<$ 0.001) membership

Weighted gene co-expression networks



Functional enrichment analysis

- Specific biological processes relevant for each candidate gene module by calculating GO terms and pathway enrichment
- Significant (p-value <0.05) GO and pathway enrichment for all modules
  - The respective Entrez gene identification was obtained through the *biomaRt* R package.
- Use the *GOstats* R package as well as the related Affymetrix Chip Expression Set annotation data to each associated organism.

Finding consensus modules

- The basic intuition, exploring co-occuring gene sets. Frequent pattern mining was first proposed by Agrawal et al. (1993)
- The consensus significance is measured by metrics like **(S)upport** and **(C)onfidence** of the co-occurring annotations

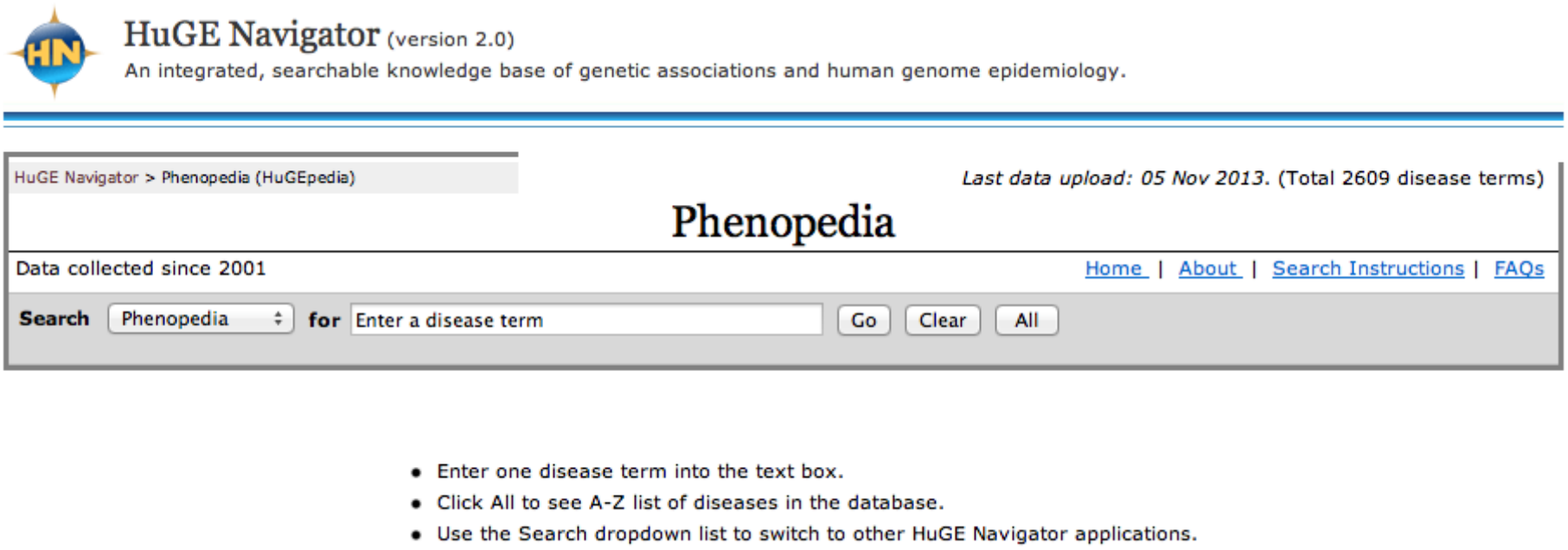
The basis of *Transactions & Itemset*

A transcriptomic study maps to a *transaction_id* and it has several gene modules

An itemset maps to an annotation (or a set of annotations)

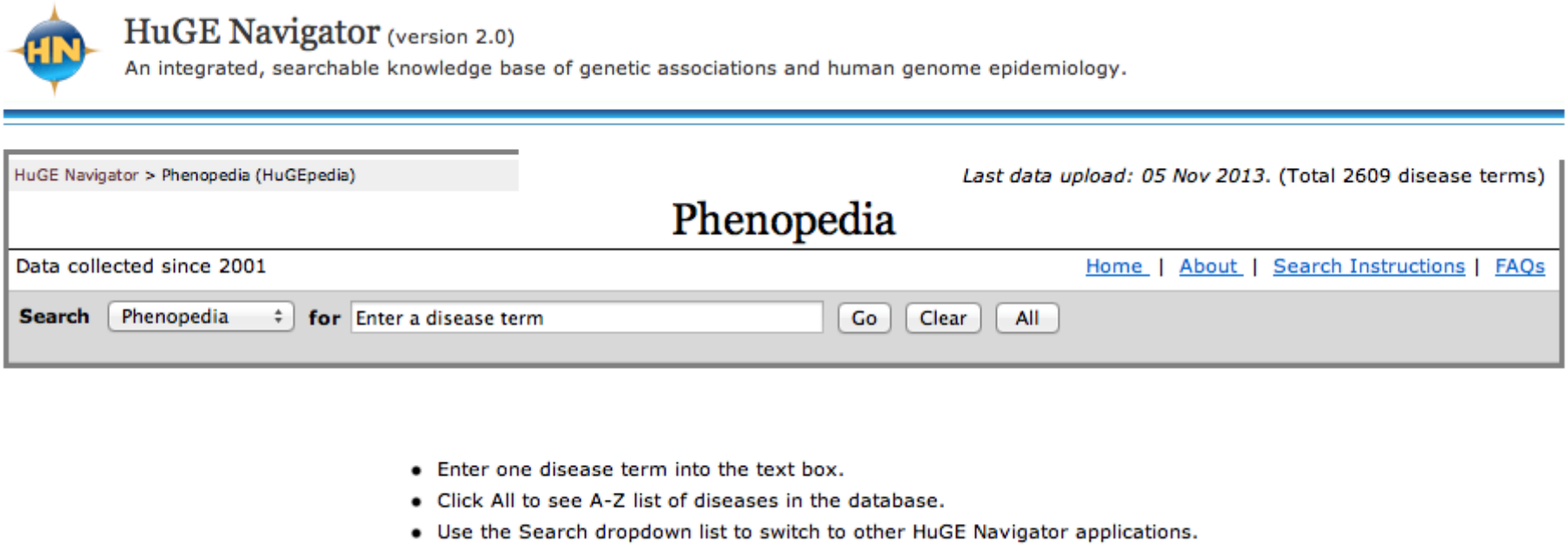
mRNA metabolic process, mRNA processing $\Rightarrow$ *RNA splicing :(S)0.3,C(1),L(2.5)*

Selection of potential T2D genes



- The gene ADIPOR1 encodes a protein which acts as a receptor for adiponectin, a hormone secreted by adipocytes which regulates fatty acid catabolism and glucose levels. Patients who developed T2D present a low activity of this gene when compared with normal ones

Selection of potential T2D genes



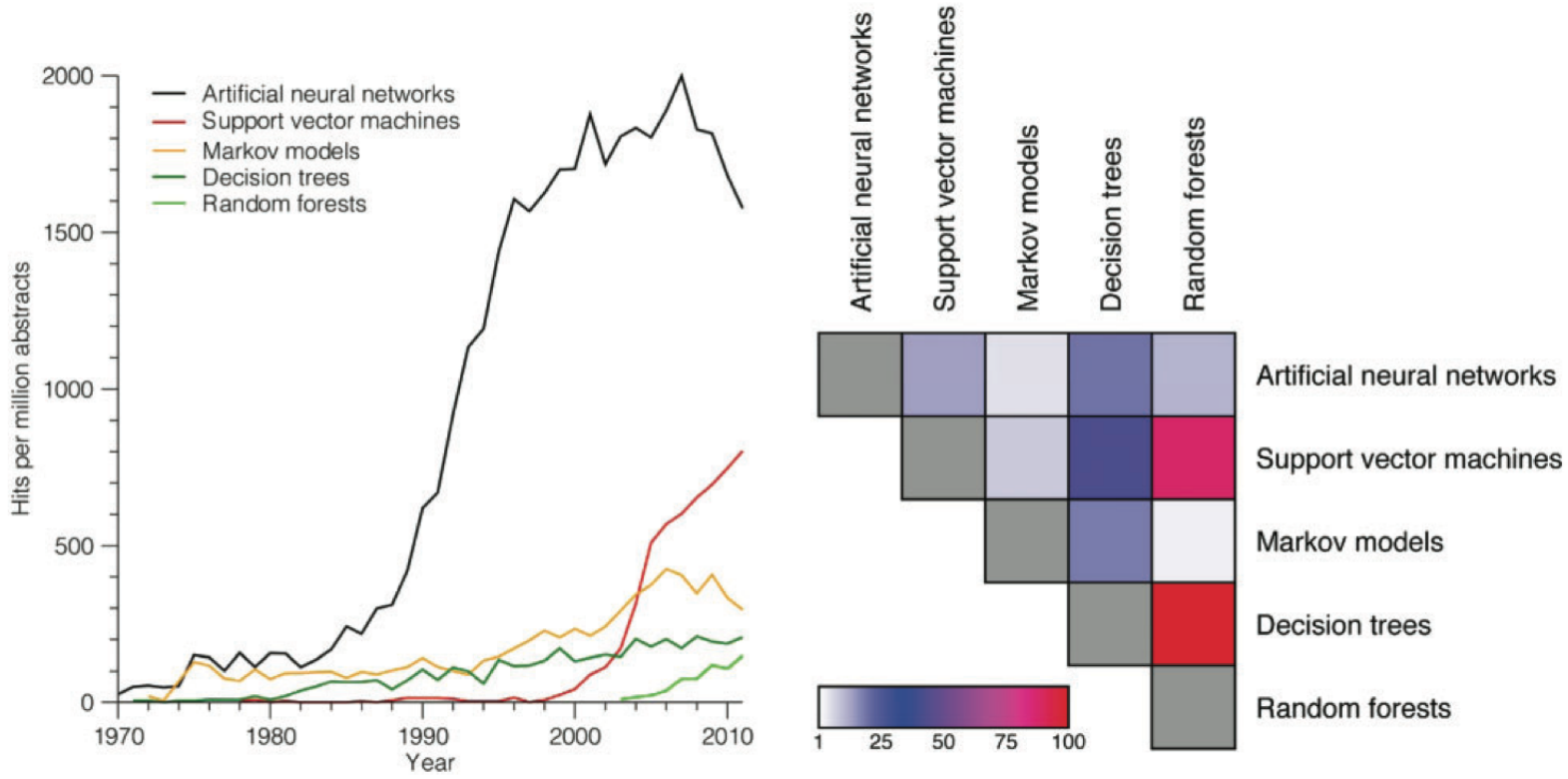
- The gene CDC123 encodes proteins highly associated to the production of insulin. Variations of this gene are also related to a low production of the hormone

Summary

- Regardless whether motif searches or network clustering is used for network decomposition, the resulting modules should not be considered as isolated components.
- Discovering functional modules
  - i) a more compact and semi-automatic way to identify gene network modules (*cliques*); ii) an optimization procedure to calculate the thresholds for the most relevant annotation across studies; and iii) extend the pipeline to deal with RNA-Seq data.
- Network semantics (nodes, edges, annotations) are essential keys to better exploration of biological network properties.

Outlook – final remarks

The rise and fall of supervised machine learning techniques



Jensen and Bateman (2011), BIOINFORMATICS Vol. 27 no. 24, pages 3331-3332

Data integration in the era of omics: current and future challenges

The data exploitation aspect of data integration is probably the one that requires most attention, as it involves:
i) the use of prior knowledge - and its **efficient storage**,
ii) the development of **statistical methods** to analyze heterogeneous data sets, and
iii) the creation of data explorative tools that incorporate both useful summary statistics and new **visualization tools**.

Gomez-Cabrero et al. (2014) BMC Systems Biology 2014, 8 (Suppl 2):I1

Data integration in the era of omics: current and future challenges

	RNA-Seq	ncRNA	ChIP-Seq Histone	ChIP-Seq TF	CpG DNA Methylation	DNase-Seq	Complete DNA sequencing	Exome sequencing	Proteomics	Metabolomics	Chromatin Conformation	Clinical Data	Co-morbidities	Other
RNA-Seq		29.6%	24.8%	29.6%	32.8%	16.0%	21.6%	22.4%	36.8%	21.6%	14.4%	28.0%	10.4%	0.0%
ncRNA	6.4%		8.0%	7.2%	10.4%	4.0%	6.4%	8.0%	5.6%	4.0%	1.6%	10.4%	4.0%	0.0%
ChIP-Seq Histone	6.4%	0.8%		16.0%	16.0%	11.2%	3.2%	4.8%	7.2%	4.0%	8.8%	5.6%	2.4%	0.0%
ChIP-Seq TF	6.4%	0.8%	0.8%		12.0%	16.0%	5.6%	7.2%	9.6%	4.0%	10.4%	7.2%	2.4%	0.0%
CpG DNA Methylation	11.2%	2.4%	3.2%	2.4%		8.8%	9.6%	7.2%	6.4%	4.0%	9.6%	12.0%	4.8%	0.0%
DNase-Seq	4.0%	0.8%	1.6%	2.4%	4.8%		4.0%	5.6%	4.8%	4.0%	10.4%	9.6%	2.4%	0.0%
Complete DNA sequencing	8.8%	1.6%	1.6%	1.6%	2.4%	4.0%		10.4%	13.6%	10.4%	2.4%	20.0%	5.6%	0.0%
Exome sequencing	17.6%	0.8%	1.6%	0.8%	2.4%	0.8%	6.4%		12.0%	8.8%	0.0%	20.0%	7.2%	0.0%
Proteomics	15.2%	1.6%	0.8%	0.8%	1.6%	2.4%	4.8%	8.0%		27.2%	5.6%	16.8%	5.6%	1.6%
Metabolomics	16.8%	2.4%	2.4%	1.6%	3.2%	2.4%	6.4%	4.8%	10.4%		2.4%	17.6%	6.4%	0.8%
Chromatin Conformation	0.8%	0.0%	2.4%	2.4%	0.8%	0.0%	0.8%	0.0%	0.0%	0.8%		4.0%	2.4%	0.0%
Clinical Data	31.2%	8.0%	7.2%	9.6%	15.2%	9.6%	20.0%	21.6%	16.8%	20.0%	4.0%		14.4%	3.2%
Co-morbidities	8.8%	4.0%	3.2%	5.6%	6.4%	4.8%	7.2%	5.6%	2.4%	5.6%	0.8%	16.0%		1.6%
Other	0.8%	0.0%	0.0%	0.0%	0.8%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%	2.4%	0.8%	

Same data Type in Basic Science	14.4%	6.4%	5.6%	6.4%	4.8%	3.2%	5.6%	4.0%	7.2%	4.8%	2.4%	4.0%	3.2%	1.6%
Same data type in Clinical Environment	5.6%	0.0%	0.0%	0.8%	0.8%	0.0%	2.4%	0.0%	1.6%	4.0%	0.0%	5.6%	0.8%	0.0%

Figure 2 Relevance of integration schemes. (a) Each matrix location (i,j) shows the percentage of survey participants that selected as relevant the integration of data type i and data type j in basic (upper matrix) and clinical (lower matrix) research. (b) shows the percentage of participants that selected as relevant the integration of the same data type for the data types included in the list.

Obrigado – Thanks – Merci!

References I

[1] Ronnie Alves, Domingo S. Rodriguez-Baena, and Jesus S. Aguilar-Ruiz. Gene association analysis: a survey of frequent pattern mining from gene expression data. *Briefings in Bioinformatics*, 11(2):210–224, 2010.

[2] Peter Flach. Machine learning: The art and science of algorithms that make sense of data. 2012.

[3] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

References II

[4] Pedro Larraaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iaki Inza, Jos A. Lozano, Rubn Armaanzas, Guzmn Santaf, Aritz Prez, and Victor Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 2006.

[5] Sanja Rogic, Alan K. Mackworth, and Francis B.F. Ouellette. Evaluation of gene-finding programs on mammalian sequences. *Genome Research*, 11(5):817–832, 2001.

[6] Sanja Rogic, B.F. Francis Ouellette, and Alan K. Mackworth. Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics*, 18(8):1034–1045, 2002.

References III

[7] Uwe Scherf, Douglas T. Ross, Mark Waltham, Lawrence H. Smith, Jae K. Lee, Lorraine Tanabe, Kurt W. Kohn, William C. Reinhold, Timothy G. Myers, Darren T. Andrews, Dominic A. Scudiero, Michael B. Eisen, Edward A. Sausville, Yves Pommier, David Botstein, Patrick O. Brown, and John N. Weinstein. A gene expression database for the molecular pharmacology of cancer. *Nat Genet*, 24(3):236–244, 03 2000.

[8] Therese Srlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Thor Thorsen, Hanne Quist, John C. Matese, Patrick O. Brown, David Botstein, Per Eystein Lnning, and Anne-Lise Brresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

References IV

[9] Laura J. van 't Veer, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, Rene Bernards, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 01 2002.